# Optimization of Geospatial Data Modelling for Crop Production by Integrating Proximal Soil Sensing and Remote Sensing Data

**Md Saifuzzaman**

**Department of Bioresource Engineering**

**Faculty of Agricultural and Environmental Sciences**

**Macdonald Campus of McGill University**

**21,111 Lakeshore Road, Ste. Anne de Bellevue, Québec H9X 3V9 Montreal, Canada**

**June 2020**

**A Thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy**

**© Md Saifuzzaman, Canada, 2020**

## Dedication

This dissertation is deeply dedicated to those who devote their efforts to the development of sustainable agricultural production systems for the future generations as well as to my wife Nyema Sultana and my children - Nabila, Nafisa and Naveed for their support, encouragement, companionship and understanding throughout the time of pursuing my doctoral degree.

# Table of Contents

# Abstract

Emerging technologies in precision agriculture (PA) offer a wide array of advanced methods to assess soil properties and to determine soil variability. Remote sensing (RS) and proximal soil sensing (PSS) technologies, widely used in quantifying surface and subsurface soil parameters, can be combined to infer spatial patterns of soil heterogeneity and to develop thematic maps for site-specific management. However, the use of these soil sensors must be reviewed constantly to maintain their efficiency and precision in delineating the soil-crop relationship and to inform PA approaches. Data mining and model optimization are key to evaluating high-density geospatial data in a dynamic production system. High-density PSS and RS-based soil characterization was explored and optimization techniques for digital soil mapping in PA were evaluated.

In a first study, sensor measurements were subjected to multivariate statistical analysis, followed by an evaluation of a new Neighborhood Search Analyst (NSA) and the capacity of other data clustering algorithms to delineate spatially contiguous zones in agricultural fields and to optimize soil sampling locations to inform best management practices. PSS-based topography, apparent electrical conductivity ($EC_a$), and RS-based indices data from 3 sites in Ontario, Canada, were employed to assess the novel technique's performance in accurate zone delineation. In creating homogeneous zones, a maximum of 70% field variance ($R^2 = 0.70$) was achieved. The $R^2$ of the k-means cluster compared to that of the NSA was relatively higher ($R^2 = 0.80$) where, the k-means cluster map consisted of groups or pixels with isolated boundaries in various parts of the field. The NSA's unique capacity, across various locations, to produce an optimum (or user-defined) number of zones highlighted its superiority to k-means' partitioning with isolated boundaries.

A second study assessed the utility of PSS-based soil characterization in developing an optimum prediction method for multiple soil properties at 12 sites across Ontario, Canada. Targeted soil sampling locations were determined and optimized using NSA clustering tools. Measured $EC_a$, topographic parameters and six lab-quantified soil properties [pH, buffer pH, soil organic matter (SOM), Phosphorus (P), Potassium (K) and Cation Exchange Capacity (CEC)] were used in evaluating the method's predictive capacity and to compare different fields' propagated soil measurement errors by drawing on the results of the North American Proficiency

Testing program. Pearson's correlation coefficients exceeding 0.60 indicated strong relationships between sensor variables and field-measured soil properties, topographic parameters and shallow $EC_a$ sensor variables, allowing effective predictions of several soil chemical properties (*i.e.*, SOM, P, and CEC).

Lastly, supervised machine learning models drawing on high-density information from multiple sensors (PSS and RS) operating at different geospatial scales, were used to generate thematic soil maps for an agricultural field in Ontario, Canada. A random forest (RF) regression model delineated the complex hierarchical relationships existing among the sensor variables and evaluated prediction efficiencies for multiple soil nutrients. The reduction of variables based on their relative importance and parameter optimization (i.e., by defining the number of trees) of the regression forest improved the predictive accuracy for nine soil properties at the cross-validation stage. The best prediction capacity has been achieved for soil pH, K, and Zn ($R^2 \geq 0.80$).

Sophisticated technologies are critical to generating finer resolution thematic maps for PA and to address soil management at various geospatial scales. Multilayer data optimization techniques used in multiple sensor-based mapping provide information of field-scale variability and soil prediction at the local-scale. This research indicated that soil variability which was determined using sensor-fused data and optimization techniques could assist in constructing precise soil property prediction models and in developing reliable thematic maps for site-specific crop management.

# Résumé

Les technologies émergentes en agriculture de précision (AP) offrent un large éventail de méthodes avancées pour évaluer les propriétés du sol et déterminer leur variabilité. Les technologies de télédétection (RS) et de détection de sol proximale (PSS), largement utilisées pour quantifier les paramètres pédologiques de surface et souterrains, peuvent être combinées de manière à déduire des modèles spatiaux d'hétérogénéité des sols et pour développer des cartes thématiques pour une gestion spécifique au site. Cependant, ces explorations avec capteurs de sol doivent être revues en permanence pour maintenir leur efficacité et leur précision dans l'encadrement des relations sol-culture et des approches PA. L'exploration de données et l'optimisation des modèles sont essentielles à évaluation des données géospatiales à haute densité dans un système de production dynamique. La caractérisation des sols à base de PSS et RS à haute densité fut explorée et les techniques d'optimisation pour la cartographie numérique des sols en PA furent évaluées.

Dans une première étude, des mesures des capteurs furent soumises à une analyse statistique multivariée, suivie d'une évaluation de la capacité de Neighbourhood Analyst (NSA) et d'autres algorithmes de regroupement de données à délimiter des zones spatialement contiguës dans les champs agricoles et d'optimiser les emplacements d'échantillonnage du sol pour éclairer les meilleures pratiques de gestion. La topographie basée sur PSS, la conductivité électrique apparente ($EC_a$) et les données d'indices basés sur RS de 3 sites en Ontario, au Canada, ont permis une évaluation des performances de la nouvelle technique dans la délimitation précise des zones. Le $R^2$ du groupe de k-moyennes par rapport à celui de la NSA était relativement plus élevé ($R^2$ = 0,80) où, la carte du groupe de k-moyennes consistait en groupes ou pixels avec des limites isolées dans diverses parties du champ. La capacité unique des NSA, sur divers sites, à produire un nombre optimal (ou défini par l'utilisateur) de zones, a mis en évidence sa supériorité sur le partitionnement par *k*-means avec des limites isolées.

Une seconde étude évalua l'utilité de la caractérisation des sols basée sur PSS dans le développement d'une méthode de prédiction optimale pour plusieurs propriétés des sols, pour 12 sites à travers l'Ontario, Canada. Les emplacements d'échantillonnage des sols ciblés furent déterminés et optimisés à l'aide d'outils de regroupement NSA. L'$EC_a$ mesurée, les paramètres topographiques et six propriétés du sol quantifiées en laboratoire (pH, pH tampon, SOM, P, K et

CEC) servirent à évaluer la capacité prédictive de la méthode et à comparer l'erreur de mesure du sol propagée de différents champs en s'appuyant sur les résultats du Programme North American Proficiency Testing. Les coefficients de corrélation de Pearson supérieurs à 0,60 indiquaient de fortes relations entre les variables du capteur et les propriétés du sol mesurées sur le terrain, les paramètres topographiques et les variables du capteur ECa peu profondes, permettant des prédictions efficaces de plusieurs propriétés chimiques du sol (c.-à-d. SOM, P et CEC).

Enfin, des modèles d'apprentissage automatique supervisé s'appuyant sur des informations à haute densité provenant de plusieurs capteurs (PSS et RS) fonctionnant à différentes échelles géospatiales ont servi à générer des cartes thématiques des sols pour un champ agricole en Ontario, au Canada. Un modèle de régression aléatoire en forêt (RF) a délimité les relations hiérarchiques complexes existant entre les variables du capteur et évalué l'efficacité de la prédiction pour plusieurs nutriments du sol. L'importance réduction en fonction de leur relative variable et l'optimisation des paramètres (c'est-à-dire en définissant le nombre d'arbres) de régression ont amélioré la précision prédictive pour neuf propriétés du sol au stade de la validation croisée. Le coefficient de détermination ($R^2$) a montré que la plus grande précision (ajustement du modèle) a été atteinte pour la prédiction du pH, du K et du Zn ($R^2 \geq 0.80$).

Des technologies sophistiquées sont essentielles à la génération de cartes thématiques à résolution plus fine pour l'AP et à la gestion des sols à différentes échelles géospatiales. Les techniques d'optimisation des données multicouches utilisées dans la cartographie basée sur plusieurs capteurs permettent de comprendre la variabilité à l'échelle du terrain et la prévision du sol à l'échelle locale. Cette recherche a indiqué que la variabilité du sol déterminée à l'aide de données fusionnées par capteur et de techniques d'optimisation pourrait aider à construire des modèles précis de prévision des propriétés du sol et à développer des cartes thématiques fiables pour la gestion des cultures spécifiques au site.

# Acknowledgements

First, I would like to wholeheartedly acknowledge all those who participated, guided and supported me in the completion of this dissertation. Without their support and guidance, it would not have been possible to combine the following elements.

This research project was supported in part by funds provided through Nature and Technology-FRQNT (B2X), Government of Quebec, Canada as well as Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) New Directions Research Program. Datasets of this research project were also supported by the Grain Farms of Ontario "Precision Agriculture Advancement for Ontario" Project with funding from Growing Forward 2 through Agricultural Adaptation Council.

I am extremely grateful to my research director Prof. Viacheslav Adamchuk, who provided me with an opportunity to work with him, for research funding, and research facilities, and expanded my knowledge and skills, and, moreover, helped me in resolving research questions and challenges over the past years. I truly admire him for his guidance, support, motivation in research and understanding throughout the years, but also for his wide range of experience and knowledge not only in research but also in daily life.

I highly appreciate the contributions given by Nicole Rabe Ontario Ministry of Agriculture, Food and Rural Affairs, Ontario) and Doug Aspinall (Woodrill Farms Ltd, Guelph, Ontario) in terms of the materials and data analyzed in this study. Special thanks to Paul Raymer and Ryan Eyre of SoilOptix® (Tavistock, ON, Canada) for gamma spectrometry mapping. I also acknowledge the valuable feedback from cooperating all farmers in Ontario.

I thank Prof. Shiv Prasher, member of my supervisory committee and the comprehensive exam committee, who constantly provided suggestions and ideas throughout this research. I also thank the other members of the comprehensive exam committee, Prof. Vijaya Raghavan, Prof. Pierre R.L. Dutilleul, and Dr. Zhiming Qi for providing recommendations to improve this research work. The efforts of Dr. David Mulla, Professor of the University of Minnesota, and Prof. Chandra Madramootoo who were the members of the dissertation examination committee for improving the quality of this dissertation are highly appreciated. I would like to thank Dr. Georges Dodds and Ms. Darlene Canning for proofreading the research articles.

I acknowledge the contributions of the past and current Precision Agriculture and Sensor Systems (PASS) research team members Dr. Nandkishor Dhawale, Dr. Wenjun Ji, and Dr. Hsin-Hui Huang, and Sophie Lauzon for their data retrieval and diagnostic work and for performing soil sampling and sensor measurements in various fields. Especially, much gratitude goes to Dr. Samson Sotocinal for DUALEM mapping and fieldwork support. Also, I am very grateful to all supportive colleagues and other PASS members and Dr. Jaesung Park, Roberto Buelvas, Eko Leksono, Amanda Boatswain Jacques, Maxime Leclerc, Marie-Christine Marmette, Dr. Salman Tabatabaia, and many others for sharing interdisciplinary ideas and being the greatest teammates and friends.

I also acknowledge the valuable contributions from members of the supervisory committee and, the co-authors of the manuscripts of this research, Prof. Shiv Prasher, Dr. Asim Biswas, Nicole Rabe, Roberto Buelvas, Doug Aspinall, and Dr. Wenjun Ji. They contributed their knowledge and expertise in reviewing my writing for publication. Their tremendous support, suggestions, and comments helped me understand my research from different perspectives.

Lastly, my sincere gratitude goes to my wife for her patience, encouragement, and support during the time of the research work, and to my children for being a source of happiness, and to my late parents for their inspiration for this degree.

## Publications, Manuscripts, Conference Papers

### a. Conference proceedings:

1. Saifuzzaman, M., & Adamchuk, V. (2017). Proximal Soil Sensing and Remote Sensing Data Processing for Precision Agriculture in Ontario, Canada. In *Abstracts from Annual Meeting of the Association of American Geographers, 5 - 9 April 2017* (pp. 1204–1205). Boston, Massachusetts, USA: (CD publication).

2. Saifuzzaman, M., & Adamchuk, V. (2017). Geospatial Analysis of Proximal Soil Sensing and Remote Sensing Data in Precision Agriculture. In *Abstracts from the Earth Observation Summit 2017, UQAM (Science Centre), 20 - 22 June 2017. Canadian Remote Sensing Society.* Montreal, Quebec, Canada: (Published on-line at https://crss-sct.ca/conferences/csrs2017).

3. Ji, W., Adamchuk, V., Lauzon, S., Su, Y., Saifuzzaman, M., & Huang, H. (2017). Pre-processing of on-the-go mapping data. In *The Book of Abstracts for Pedometrics 2017 Conference, 26 June-1 July 2017* (p. 113). Wageningen, the Netherlands.

4. Saifuzzaman, M., Adamchuk, V., Huang, H., & Biswas, A. (2018). Integration of Proximal Soil Sensing and Remote Sensing Data in Agriculture. In *Abstracts from the 39th Canadian Symposium on Remote Sensing 2018, University of Saskatchewan, 19 - 21 June 2018.* Saskatoon, Canada: (Published on-line at https://crss-sct.ca/conferences/csrs-2018).

5. Saifuzzaman, M., Adamchuk, V., Huang, H., Ji, W., Rabe, N., & Biswas, A. (2018). Data Clustering Tools for Understanding Spatial Heterogeneity in Crop Production by Integrating Proximal Soil Sensing and Remote Sensing Data. In *Proceedings of the 14th International Conference on Precision Agriculture, 24 - 27 June 2018. International Society of Precision Agriculture* (p. 14). Montreal, Quebec, Canada: (Published on-line at http://www.ispag.org).

6. Saifuzzaman, M., Adamchuk, V., Biswas, A. & Dutilleul, P. R. L. (2019). Soil Prediction using High-Density Data for Understanding Field Variability and Crop Management. In *Abstracts from Annual Meeting of the Association of American Geographers, April 3 - 7 2019*. Washington DC, USA: (CD publication).

7. Saifuzzaman, M., Adamchuk, V., Biswas, A., Prasher, S., & Rabe, N., (2019). Geospatial Data Modelling by Integrating Sensor-Fused Data in Agricultural Field Management. In *Proceedings of the 13th Pedometrics conference 'Pedometrics 2019, June 2 - 6.* Guelph ON, Canada: (Published on-line at http://www.pedometrics2019.com).

8. Saifuzzaman, M., & Adamchuk, V. (2020). Optimization of Geospatial Data Modelling for Crop Production by Integrating Proximal Soil Sensing and Remote Sensing Data. In *Proceedings of the 15th International Conference on Precision Agriculture, 28 June - 01 July 2020. International Society of Precision Agriculture* (p. 1). Minneapolis, Minnesota, USA: (Accepted & published on-line at http://www.ispag.org).

9. Saifuzzaman, M., Adamchuk, V., & Rabe N. (2020). Sensor-Fusion by Machine Learning Methods for Field-Scale Thematic Soil Mapping. In Abstracts from the 41st Canadian Symposium on Remote Sensing 2020, University of Lethbridge, 13 - 16 July 2020. Lethbridge AB, Canada: (Published on-line at https://crss-sct.ca/conferences/csrs-2020).

**b. Journal publications:**

1. Saifuzzaman, M., Adamchuk, V., Buelvas, R., Biswas, A., Prasher, S., Rabe, N., Aspinall, D., & Ji, W. (2019). Clustering Tools for Integration of Satellite Remote Sensing Imagery and Proximal Soil Sensing Data. *Remote Sensing – MDPI* 11(9).

2. Saifuzzaman, M., Adamchuk, V., Biswas, A., and Rabe, N. (2020). High-density Proximal Soil Sensing Data and Topographic Derivatives to Characterize Field Variability. *Biosystems Engineering - Elsevier* (In preparation).

3. Saifuzzaman M., Adamchuk, V., and Biswas, A. (2020). Optimization of Random Forest Model for Sensor Data Fusion and Thematic Soil Mapping at Field Scale. *Remote Sensing-MDPI* (In preparation).

4. Saifuzzaman M., Adamchuk, V., Biswas, A., and Prasher, S. (2020). Remote sensing and Proximal Soil Sensor Data Fusion using Geospatial Model for Mapping Agricultural Fields. *Remote Sensing of Environment- Elsevier* (In preparation).

## Contributions of Authors

The chapters of this thesis have been presented at scientific conferences, published, and submitted for peer-reviewed journal publications. This thesis consists of four manuscripts, of which the author was fully responsible for developing the research questions, data clustering, and model design, analytical approaches including programming and data exploration, interpreting results, and writing this dissertation and the manuscripts. However, this thesis would not be possible without the contribution of Prof. Viacheslav Adamchuk, the supervisor of this dissertation and the co-author of all manuscripts, who provided conceptual frameworks, scientific guidance, and advice in the development and reviewing of these manuscripts.

Besides the research design and compilation of the thesis, Chapter 3 was published in *Remote Sensing (MDPI)* journal, in the section "Remote Sensing in Agriculture and Vegetation". The coauthors provided substantial contributions in terms of technical edits (Dr. Asim Biswas), in designing the algorithm (Roberto Buelvas), in remote sensing analysis (Prof. Shiv Prasher), in sharing geospatial data for three sites in Ontario (Nicole Rabe), and for advice on study sites (Doug Aspinall). Chapter 4 will be submitted to *Biosystems Engineering (Elsevier)* after addressing the coauthor's valuable comments and technical advice (Dr. Asim Biswas). In this context, we also thank Nicole Rabe for various geospatial data support at thirteen study sites in Ontario. We plan to submit Chapter 5 to *Remote Sensing (MDPI)*, for the special issue "Estimation and Mapping of Soil Properties Based on Multi-Source Data Fusion." This paper is co-authored with Dr. Asim Biswas, who provided advice on soil science and technical issues. We plan to submit a summary paper with field validation (Chapter 3 and Chapter 5) to *Remote Sensing of Environment (Elsevier)*, to be co-authored with Dr. Asim Biswas and Prof. Shiv Prasher, who provided technical advice and comments for improvement of the paper. Besides the data collection activities for the research papers, many research students and teaching staff in the Bioresource Engineering Department, Macdonald Campus of McGill University were helpful in reading and revising the manuscripts.

## List of Tables

## List of Figures

## List of Abbreviations and Symbols

$a$ – upland contributing area

AR – aspect ratio

$\beta$ – slope in TWI calculation

BpH – Buffer pH

Ca – Calcium

CART – classification and regression tree

CEC – cation exchange capacity

CEC – cation exchange capacity

CV – cross validation

DAE – delineated areal extent

DEM – digital elevation model

DSM – digital soil mapping

DVI – Difference Vegetation Index

$EC_a$ – soil apparent electrical conductivity

$EC_a^{0 \cdot x}$ – $EC_a$ from surface to depth $x$ in meters

FPI – fuzziness performance index

GIS – Geographic Information System

GNSS – global navigation satellite system

GNSS – global navigation satellite systems

GPS – Global Positioning System

HCP – horizontal coplanar

$k$ – individual bootstrap sample

K – potassium

$K$ – set of regression trees

MAD – median absolute deviation

Mg – Magnesium

Mn – Manganese

MSE – mean square error

MZA – management zone analyst

N – total number of grid cells

NAPT – North American Soil Proficiency Testing program

NCE – normalized classification entropy

NDRE – Normalized Difference Red Edge Index

NDVI – Normalized Difference Vegetation Index

NIR – Near infrared spectrum

NSA – Neighborhood Search Analyst

OF – objective function

OK – ordinary kriging

OLS – ordinary least square

OMAFRA – Ontario Ministry of Agriculture Food and Rural Affairs

OMNRF – Ontario Ministry of Natural Resources and Forestry

OOB – Out of bag

P – phosphorus

PA – precision agriculture

PRP – perpendicular coplanar

PSS – proximal soil sensing

$r$ – Pearson's correlations

$R^2$ – coefficient of determination

$R^2_{adj}$ – adjusted r-squared

RE – red edge

RED – visible red reflectance

RF – random forest

RMSE – root mean square error

RPD – performance to deviation

RS – remote sensing

RTK – Real-Time Kinematic

SCZ – spatially contiguous zones

STD – standard deviation

SE – standard error

SOM – soil organic matter

TC – Total count

$T_k$ – the individual learner or decision tree

TWI – topographic wetness index

VIS (or RGB) – Visible spectrum

VRA – variable rate application

WAAS – wide area augmentation system (GPS)

Zn – Zinc

$^{232}$Th – Thorium 232 isotope

$^{238}$U – Uranium 238 isotope

$^{40}$K – Potassium 40 isotope

$\overline{y_m}$ – mean measured value

$y_{m_i}$ – $i^{\text{th}}$ measured value

$y_{p_i}$ – $i^{\text{th}}$ predicted value

# Chapter 1: Introduction

## 1.1 General introduction

Recently, global agricultural production systems have faced various challenges due to environmental degradation. The stresses of the production environment are determined by the intensification of production and poor management. This issue highlights the fact that feeding the world's growing population necessitates improved production systems, particularly through the use of precision agriculture technologies (Oliver *et al.*, 2013). To better understand such a production environment, one must begin by defining Precision Agriculture (PA) and its components, including spatiotemporal variability (*i.e.*, in the field, soil, crop), farm management, profitability, and environmental sustainability (ISPA, 2019). In an ongoing effort to improve crop production systems, agricultural scientists have focused on these sectors in assessing soil-crop-environment relationships and managing system variability. Accordingly, robust technologies applied in an efficient manner can increase production quality and maximize the farm's profitability.

To best circumscribe the above issues, a comprehensive definition of Precision Agriculture (PA), adopted by International Society for Precision Agriculture (ISPA) in 2019, is given below:

*"Precision Agriculture is a management strategy that gathers, processes and analyzes temporal, spatial and individual data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production."*

In the last decade, three interlinked production phenomena have been considered in efforts to improve agricultural production systems and monitor their soil-environment relationship:

(i) static conditions of a production field, including the characteristics of soil and crop yields;

(ii) field management practices, mainly the production system's inputs and outputs, *e.g.*, tillage, soil amendments, crop types/rotations, harvesting, etc., and

(iii) dynamic conditions of the production environment, *e.g.*, climatic conditions, economic situation, and profitability.

These three elements are key to a farm's long-term productivity and economics which ultimately benefit the farming community and result environmental sustainability.

The static condition of the production system entails the production field's soil nutritional elements and crop yield. Soil thematic maps can provide the precision agriculture farming community substantial information on field variations, allowing for an improved farming environment according to site-specific requirements. A key challenge in preparing soil thematic maps lies in determining how to accurately represent the spatiotemporal heterogeneity of agricultural fields at different scales. Various technologies assist agricultural researchers studying precision farming systems to improve local- and regional-scale thematic maps (Zhang *et.al.,* 2002), thereby, providing a range of solutions to improve the production environment and increase its profitability. These researchers also assess field variability through their understanding of field dimensions, topographic characteristics and historical cropping records (Grunwald, 2015). Seeking to improve thematic maps and develop sustainable production strategies, this sphere of research explores the spatiotemporal dynamics of soil within the production environment.

A Digital soil map (DSM) refers to a quantitative soil prediction criterion on a geographically unique space, which draws upon dense soil and environmental covariates to infer spatiotemporal variations within the map (McBratney *et al.,* 2003; Guo *et al.,* 2015). By reducing environmental impacts, it also alters the productivity and sustainability of the landscape (Oliver *et.al.,* 2013). While exploring soil-environment dynamics using soil maps is a step forward in managing crop production systems, the process remains a black box. The implementation of DSM implies the use of several modeling and mapping techniques to better understand the spatial and temporal dynamics in different situations. However, a DSM can present several drawbacks when applied to dynamic crop growing conditions. When site-specific optimization of an agricultural field is based on inaccurate information and draws upon expensive laboratory soil analyses, economic losses may occur, and environmental concerns develop. These issues account for the relatively slow adoption of site-specific crop management in many regions.

In preparing DSMs and applying them to site-specific management, various proximal soil sensing (PSS) and remote sensing (RS) platforms have been used to gather affordable high-density multivariate datasets and explore various data indices. Most sensors measure soil parameters indirectly rather than directly, and then infer agronomic indicators of the crop growing

environment. Due to various aspects and scales in the data-mining protocol, intensive data processing and integration are required to produce maps which provide detailed characterization of field heterogeneity. Models for spatial prediction incorporate different advanced algorithms and approaches, such as geostatistics, advanced regression and machine learning. Due to the increased applicability of density maps at the local or regional scale, DSM requires optimization through modelling and validation of its performance (Minasny and McBratney, 2016).

The present research proposed several approaches to further improve digital soil modeling and thematic mapping. To better understand an agricultural field's spatiotemporal variability by way of multiple sensor information, a data clustering model must initially be implemented at various scales to assess the diverse environmental dynamics and field optimization conditions. Moreover, a heightened level of confidence, developed through error optimization and validation, improves the capacity for prediction in generating accurate soil variability maps and ultimately, in assessing soil health. Finally, machine learning algorithms (model) are necessary to evaluate the prediction results and determine whether the production system under study is economically viable. Ultimately, all such efforts assist in improving the soil and production systems and enhancing environmental sustainability initiatives. These efforts will also empower growers who operate under different management conditions and produce either cash or specialty crops, fruits and vegetables, or nursery crops — as well as their advisors (e.g., agronomists, extension agents) — to adopt precision farming.

### 1.2 Problem statement and rationale

Recently, farmers have benefited from the development of soil sensors and their use in understanding pedogenic processes and the level of soil constituents (Brown, 2006). Also, they have extensively employed various geospatial technologies and tools, such as proximal soil and remote sensors, along with Global Positioning System (GPS) and Geographic Information System (GIS), to integrate agricultural farm management and decision support systems for precision agriculture (Franzen and Mulla, 2015). However, there remain various issues with multivariate and dense soil response, data modeling, and optimization. Those sensing technologies, along with advanced methods, are used to obtain soil property information and determine soil variability for precision agriculture.

Proximal soil sensors, topsoil images, geospatial analysis, and decision support tools are essential to evaluating soil types and their variability. However, if erroneous data is combined with outdated measurements and approaches, this results in support system inaccuracies, leading to imprecise or inaccurate predictions which delay the decision-making process. Accordingly, one needs to minimize errors and render precise field assessments when using precision technologies. To better manage topsoil and improve crop production, agronomists and production farmers need to periodically assess soil fertility in a cost-effective manner that improves upon conventional procedures. This research explores the optimization of geospatial data modeling and thematic mapping strategies by assessing multivariate PSS measurements obtained through satellite imaging data and conventional laboratory measurements of soil parameters from different parts of a single field or from different agricultural fields.

Soil horizon and its boundaries have important effects on the physiochemical properties of topsoil, and are the major issues of contention in assessing soil health and sustainability (Oliver *et al.*, 2013). In the soil boundary delineation process, topography is considered the variable having the greatest influence on soil constituents, influencing both soil physical and chemical properties which are key to agricultural production. In the mapping process, crop vigor and historical vegetation trends along with topographic characteristics are assessed as indicators of soil health and farm management. However, the key challenges are to identify the relative importance of each variable and to optimize the selected variables at different spatial scales. The optimized data can then be recommended for use in rigorous modelling procedures.

Traditional data, conventional methods and thinking often make the management of soil variability time consuming and costly in small and local-scale agricultural fields. Recent research has explored technologies and quantitative methods that make it possible to infer the spatial pattern of soil heterogeneity for digital soil mapping. The soil response to crop yields and its inherent properties are precisely evaluated through numeric prediction modeling. Soil-landscape modeling using supervised or unsupervised methods is needed to better understand geospatial variability and achieve greater accuracy (Grunwald, 2006). At present, the key research question is how multivariate data at different scales can be integrated in such a manner as to integrate multiscale variability into spatiotemporal assessments under crop production. As part of this initiative, this research examines high-density data mining techniques and various sensor data fusion algorithms

and optimization techniques to predict the heterogeneity of agricultural landscapes for making the thematic soil map.

Many sensor-measured variables are linked and modeled with measured soil properties to provide a better understanding of the soil profile and nutrient content. This research generates knowledge of the spatial distribution of soil attributes, allows the production of thematic soil variability maps and provides a guideline on managing the soil quality of a farm. After a comprehensive soil assessment through modeling and the generation of thematic maps, crop advisors and farmers can rely on location-based information and crop-specific nutrient requirements to make agronomic decisions. This research will also promote the adoption of sustainable agricultural production systems by farmers, thereby, optimizing zonal agricultural inputs and ultimately leading to the adoption of best management practices.

### 1.3 Research objectives

The overarching goal of this research is sensor-based soil characterization that leads to digital soil mapping. The purpose of this project was to integrate sensor data from multiple sources and to evaluate the optimization techniques and assess their usefulness in predicting different soil properties. This was done by assessing geospatial data modeling and assigning calibration zones to soil properties for various management issues. This research will generate methods of developing a prediction model and thematic soil maps for examining soil health in crop production and agricultural farm management. These goals will be accomplished through the following specific objectives:

1. implementing a Neighbourhood Search Analysis algorithm using an open-source programming platform to enable hierarchical spatial clustering of high-density and multilayer information evaluating agricultural fields.
2. assess proximal soil-sensing-based predictability of soil attributes for a series of agricultural fields under different agro-climatic conditions.
3. explore the potential for integrating proximal soil sensing data with remote sensing imagery and models to delineate field variability which is then suitable for differentiated management decisions.

## 1.4 Thesis organization

This thesis consists of **seven chapters** and covers the three objectives in detail. In **Chapter 1**, the research is introduced, and an overview of agricultural research, scope of the work, problem statement and objectives pertaining to the research question are provided. In **Chapter 2**, proximal and remote sensing-based data characterization are reviewed and their implications for the development of geostatistical and ensemble machine learning frameworks for geospatial prediction, modeling and thematic soil mapping are discussed. In **Chapter 3**, high density multivariate field characterization data is discussed, and the matter of data variability addressed. This involves the use of an unsupervised clustering analysis algorithm using multiple layers of geospatial, proximal- and remote-sensing, as well as data integration. Accordingly, hierarchical and multivariate data clustering tools are compared to traditional clustering methods for soil mapping. **Chapter 4** deals with various aspects of proximal soil sensor (PSS) data through an assessment of data quality and soil property prediction capability. This section also covers PSS data analysis methods and uncertainty analysis for model building and spatiotemporal soil mapping. In addressing the final objective, **Chapter 5** presents a supervised learning algorithm that integrates PSS and RS data along with field measurements for precise prediction and thematic soil mapping. This multivariate geostatistical model is assessed based on a regression method of different observed parameters at different stages, and its behavior in digital soil mapping. **Chapter 6** includes a tangible summary and conclusions of this research. Finally, **Chapter 7** presents contributions to knowledge, followed by recommendations and suggestions for future studies.

# Chapter 2: Review of Literature

## 2.1 Improvements in data clustering for identifying field heterogeneity and zones of soil homogeneity

Management zone delineation using different sensor data has become important in assessing soil health and crop production (Fridgen *et al.*, 2004; Vrindts *et al.*, 2005; Li *et al.*, 2007). To generate a map of the variability of management zones for application in precision agriculture, sensor data clustering tools are used to analyze and assess information regarding soil properties and to determine soil variability (Shatar and McBratney, 2001; Fridgen *et al.*, 2004; Dhawale *et al.*, 2014; Albornoz *et al.*, 2018). Cluster maps developed for monitoring site selection are valued by agronomists and extension agents for their role in informing the agronomic and management decisions they make (Adamchuk *et al.*, 2011). Recent research illustrates how cluster maps and homogeneous zones have been used for targeted sampling and optimized designs in zone specific locations in an agricultural field (Dhawale *et al.*, 2016; Albornoz *et al.* 2018). However, to attain hierarchical clustering tools which provide greater benefits in agricultural applications, their efficiency in describing the variability of different agricultural fields must be assessed in comparison with traditional clustering methods.

In the zone delineation process, high-density and high-resolution data from proximal soil sensing (PSS) and remote sensing (RS) technologies are used to infer the spatial pattern of soil heterogeneity (Deng *et al.*, 2003; Adamchuk *et al.*, 2004; Cohen *et al.*, 2013; De Benedetto *et al.*, 2013). Unsupervised methods have been widely used to assess spatial variability of high-density data and to determine a solution by isolating homogeneous field areas and potential management zones (Vrindts *et al.*, 2005; Li *et al.*, 2007; Cressie and Kang, 2010; Adamchuk *et al.*, 2011, Dhawale *et al.*, 2016). The use of unsupervised methods guided by multivariate data clustering techniques is imperative to achieving significant benefits from identifying and understanding soil variability within a production field (Burrough *et al.*, 1997; Ruß and Brenning, 2010).

Non-hierarchical cluster analysis through fuzzy logic (c-means or k-means), a form of an unsupervised model (Vendrusculo and Kaleita, 2011), is used extensively for agricultural data mining (De Gruijter *et al.*, 1997; Bragato, 2004; Gui-Fen *et al.*, 2007; Panda *et al.*, 2012). Due to

the imprecision and limitations in the isolation process of fuzzy logic (Johnson, 1967; 10. Arabie, and Hubert, 1996; Burrough *et al.*, 1997; Albornoz *et al.*, 2018), recent studies recommend using multivariate clustering tools along with hierarchical methods to represent unique thematic maps and zonal boundaries based on the homogeneity of the agricultural field (Figure 2.1) (Ruß and Kruse, 2011; Dhawale *et al.*, 2016). Based on the potential benefits of the method, Castrignanò *et al.* (2017) and others (Schueller, 2010; Dhawale *et al.*, 2014; Córdoba *et al.*, 2016; Saifuzzaman *et al.*, 2018) have proposed sensor data fusion and geostatistical approaches for building homogeneous zones. However, most of the clustering algorithms used in zone delineation inadequately handle high-density data files with multiple variables and have limited accessibility (Berkhin 2006; Viscarra Rossel *et al.*, 2011; Córdoba *et al.*, 2016). As a result, agricultural scientists and farmers often experience difficulties with variable rate operations due to fragmented management zones, which is what this clustering technique often generates (Albornoz *et al.*, 2018). Moreover, these tools often fail to fulfill current demands given their lack of validation datasets from these zones. To counter this, new, open source, enhanced clustering techniques are needed.

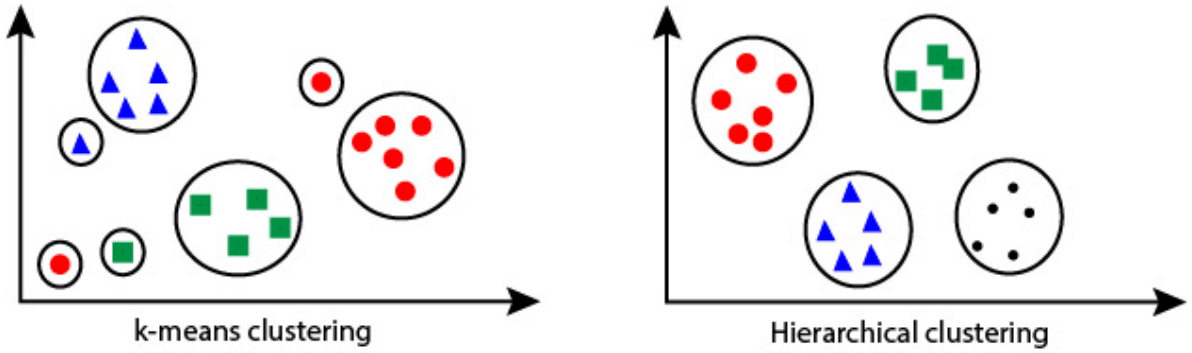



**Figure 2.1** Traditional k-means clustering method showing zones with various isolated pixels, whereas hierarchical clustering method showing well-defined zones for understanding field variability**.**

Traditional soil sampling, followed by laboratory analysis, is time-consuming, labor intensive and costly (Ji *et al.*, 2019). Adamchuk *et al.* (2011) proposed targeted sampling methods and hierarchical data analysis tools to process and manage agricultural soil sensor data. As an alternative to expensive licensed tools, open source data clustering algorithms promise new hope for large farms where sampling sites can be targeted and optimized. Besides the different clustering approaches which are available in open source libraries, (*i.e.*, R and C packages) (Albornoz *et al.*, 2018), the Python system supports data analysis libraries that are easy to use, versatile, and well-

supported. However, this open source system must be calibrated against other methods, and zone delineation must be validated to ensure the stability and precision of management decisions.

## 2.2 Use of proximal soil sensor data to estimate soil nutrients and delineate soil heterogeneity in precision agriculture

In the age of precision farming, the research scientist is a critical link in assessing crop nutritional requirements and their distribution patterns (Adamchuk *et al.*, 2004; Alchanatis and Cohen, 2013). These requirements are assessed predominantly through dense subsoil information (Lück *et al.*, 2009). To fulfill the current demand, agricultural technologies are being developed with the guidance of PSS and RS technologies (Alchanatis and Cohen, 2013; Viscarra Rossel and Adamchuk, 2013). Due to the significant data processing time and concerns about local-scale precision, large grain producers rely on spatial and temporal topsoil and subsurface information (Zhang *et al.*, 2002; Kerry *et al.*, 2017). Proximal sensing systems are an effective method for collecting density information for agricultural research (Viscarra Rossel *et al.*, 2011), and have served as a non-invasive procedure for producing fine-scale topsoil characteristics for experimental or local farms. New PSS technologies (gamma-ray spectrometry, apparent electrical conductivity) have been used to obtain high-density data, revealing the spatial distribution of edaphic properties across agricultural fields (Figure 2.2).

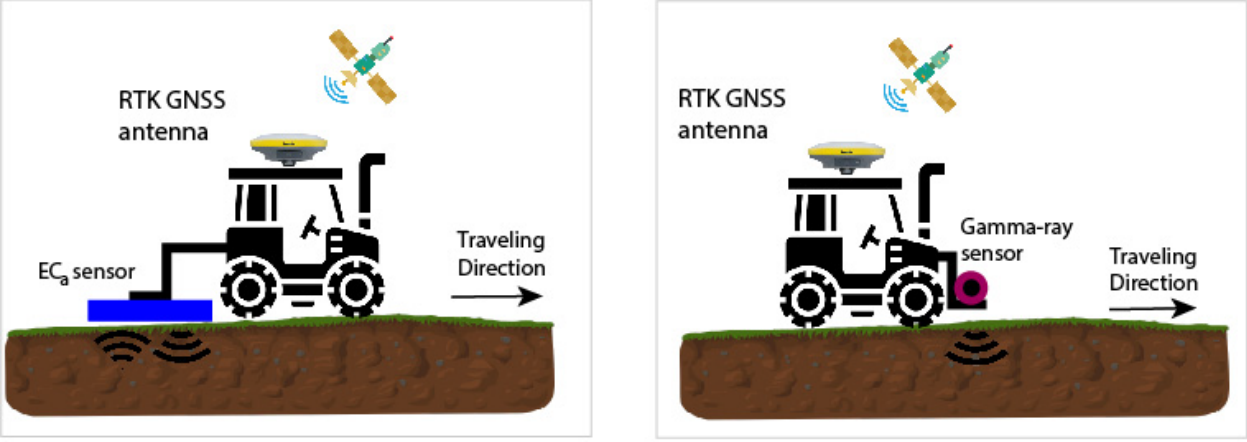



**Figure 2.2** Proximal soil sensors: Active (apparent electrical conductivity) and passive (gamma-ray) systems, a non-invasive procedure, for high-density soil mapping.

Over the past decades, passive and active PSS systems have been used to understand the soil-topography relationship and assess spatial variability for precision farming (Brown, 2006). Data from geoelectrical and electromagnetic sensors are widely used to identify soil dielectric properties and geospatial variability (Adamchuk and Viscarra Rossel, 2010). Used primarily and widely to dictate soil management in precision farming, electromagnetic energy-enabled PSS sensors (e.g., DUALEM-21S, EM-38 etc.) have been employed to provide apparent electrical conductivity ($EC_a$) measurements at variable depths to inform soil management practices under precision farming. Corwin and Lesch (2005) and Friedman (2005) found $EC_a$ measurements to directly correlate with top- and sub-soil physical properties, such as depth of clay layer, soil salinity, and soil water content, etc. Due to the variable density of the sensor measurements across soil depths, $EC_a$ data requires further calibration for site-specific depth exploration before linking it to soil constituents (Sun *et al.*, 2011; Zare *et al.*, 2018).

High-density sensor measurements are important for making agronomic decisions in precision agriculture (Appendix A). Widely implemented on a single platform, Real-time kinematic (RTK) global navigation satellite systems (GNSS) are combined with other sensors to construct dense georeferenced maps of surface topography that correspond with other measurements. A digital terrain model is generated and the topographic derivatives then are used in predicting soil attributes mapping (Bishop and Minasny, 2006). Many terrain model derivatives [*e.g.*, topographic wetness index (TWI), slope, and aspect etc.] are able to assess topographic variability, water movement, and water holding capacity for crop growth (Odeha *et al.*, 1994). Along with topographic variables, the $EC_a$ measurements are also used for predicting the presence and states of primary and secondary soil nutrients (Taylor *et al.* 2003; Adamchuk and Viscarra Rossel, 2011; Dao, 2017). The georeferenced locations, lab-measured soil analysis data, and other corresponding sensor measurements can then be used to make management decisions in agricultural fields. Adamchuk and Viscarra Rossel (2010) and Hengl *et al.*, (2017) concluded that the analysis of variables from geospatial and geostatistical data supported a predictive approach, and were valuable for the calibration of management tools in precision farming.

Geospatial analyses of different sensor variables are key in developing measurements tools employed in precision agriculture (Adamchuk and Viscarra Rossel, 2011; Hengl *et al.*, 2017). Using dense georeferenced measurements to achieve an authentic solution involves data

processing tools, approaches and models. Improved geostatistical data analysis and prediction methods are used to manage sensor measurements and soil prediction (Taylor *et al.*, 2003). Sun *et al.* (2011) and Viscarra Rossel *et al.* (2011) found that using a variable data structure for the different PSS measurements improved the accuracy of prediction for soil properties, and thereby, provided additional information in thematic mapping. The relationship among different available sensor variables are important in data mining and decision-making processes. Multivariate statistical methods (*e.g.*, correlation and regression, principle component analysis, and semi-variogram) are commonly utilized for data preprocessing and structural analysis (McBratney and Pringle, 1999; Córdoba *et al.*, 2013). Accordingly, multivariate regression analysis and prediction modeling have become popular approaches for soil characterization and the prediction of macro- or micro-nutrients.

Uncertainty analysis of the prediction model is an emerging challenge in precision soil mapping (Bishop and Minasny, 2006). To quantify the model accuracy, various statistical tests are performed and compared to the mean squared error (MSE) values of the validation points. Accordingly, D-optimality criteria and Latin hypercube sampling (LHS) have been used as model validation techniques (Adamchuk *et al.*, 2011; Panayi *et al.*, 2017). In previous studies, model sensitivity and errors are reported by different methods and minimized through different procedures [*e.g.*, ratio of performance to deviation (RPD), standard error (SE) of estimation, standard deviation (STD) of the sample, coefficient of determination ($R^2$) etc.] (Oliver, 2010; Minasny and McBratney, 2013; Sudduth, *et al.*, 2013). Thus, in the present study, the ratio of the SE of prediction to the STD of the sample serves to assess the model's performance. Moreover, to explain the proportion of variation in the regression line of the estimates, the modified version of the coefficient of determination (adjusted $R^2$) is assessed. Models with minimum propagated errors are recommended for thematic mapping and soil management in precision agriculture.

With rigorous data processing and analysis, models may be used to assess spatiotemporal variations due to annual crop nutrient uptake and amendment requirements. Hence, agricultural scientists have proposed different prescription maps for production fields based on soil variability. Recent developments in authenticated modeling requirements may integrate high-density PSS and RS data to identify comprehensive soil elements and their horizontal distribution in field-scale mapping and precision farming (Zhou *et al.*, 2016; Castrignanò et al, 2017).

## 2.3 Use of remote sensing images to delineate geospatial variability in soil and crop mapping

Various altitude and multispectral remote sensing technologies are an established non-destructive method to gather information to direct agricultural crop management (Figure 2.3). Various methods have been used to manage large farms and decision-making processes for regional scale soil improvement (Hatfield *et al.*, 2008; Salama, 2011; Rodriguez-Moreno *et al.*, 2017). In the past decades, potential challenges to data analysis in this application were coarser spatial resolution, longer revisit periods, and costly data for site-specific management (Xue and Su, 2017). Substantial improvements in spatial resolution (0.30 to 0.50 m), temporal resolution (1-3 days), and spectral resolution (3 to 250 bands) have been made accessible in recent decades (Figure 2.4), thereby, enhancing agricultural applications (Mulla, 2013; Borgogno-Mondino *et al.*, 2018). The current effort in mapping soils and developing site-specific crop management is focused on synchronizing data from low-attitude remote sensing (*i.e.*, UAV and an aerial camera with a high spatial resolution) and freely available high-attitude satellite (with medium resolution) (Mulla, 2013). The resulting datasets and derived indices have proven useful for mapping and predicting soil characteristics, such as soil moisture, organic matter, soil texture, clay content, and pH (Gregory *et al.*, 2006; Xu *et al.*, 2018).

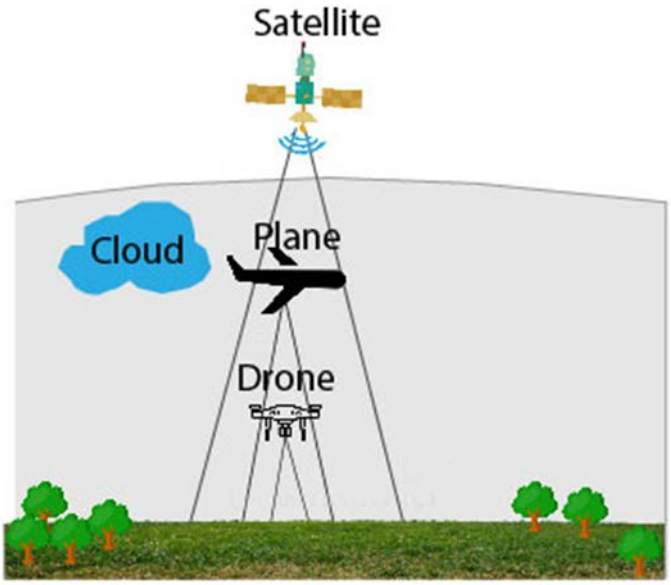



**Figure 2.3** Images taken from various altitudes and platforms for agricultural field management.

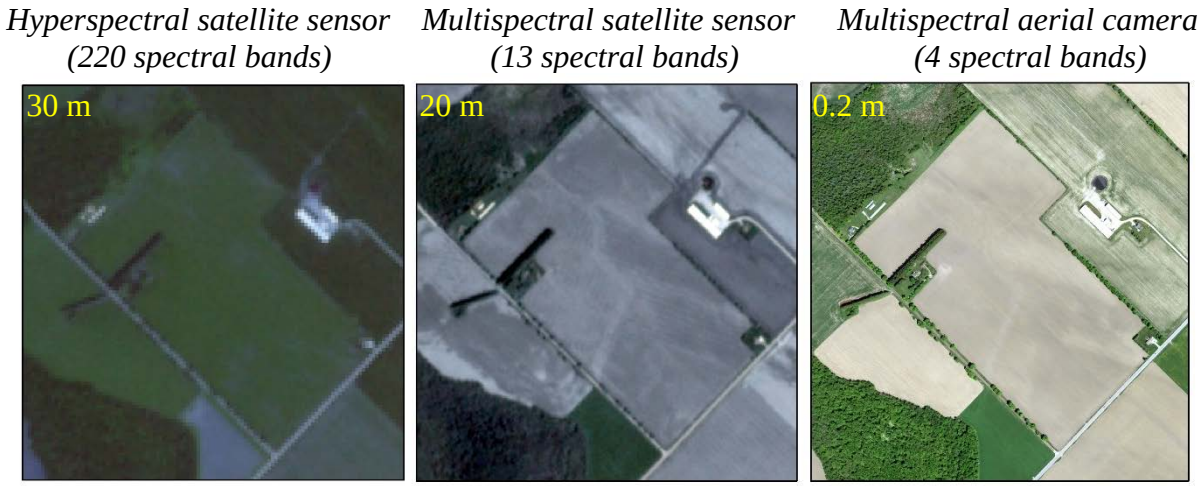



**Figure 2.4** Different spectral images, showing surface reflectance, are available for agricultural field management.

Plant health and crop growth are assessed using different remote sensing data, image indices, and data integration approaches from various platforms (*e.g.*, Sentinel-2, Rapid-Eye, and other higher resolution satellites) (Viña *et al.*, 2011). Besides, Mulla (2013) and Wulf *et al.* (2015) have indicated the availability of temporal images, highlighted the importance of nearly real-time data for crop growth assessment, and discussed the usefulness of image indices in crop and pest management (Hatfield *et al.*, 2008). A variety of spectral indices are calculated from near-infrared (NIR) and Red reflectance bands to find the ground surface's best-fit line and assess vegetation biomass (Figure 2.5 and 2.6). The Normalized Difference Vegetation Index (NDVI) and Soil Adjusted Vegetation Index (SAVI) have become the generic indices for the comprehensive assessment of crop health in agricultural communities (Mulla, 2013; Xue and Su, 2017). Likewise, the Normalized Difference Red Edge Index (NDRE) derived from near-infrared and red edge spectrum are used to detect small changes in vegetation canopy. Gitelson *et al.* (2003) illustrated that using the vegetation fraction (%) for qualitative measurement of chlorophyll content and vegetation health provided greater accuracy than simple NDVI indices.

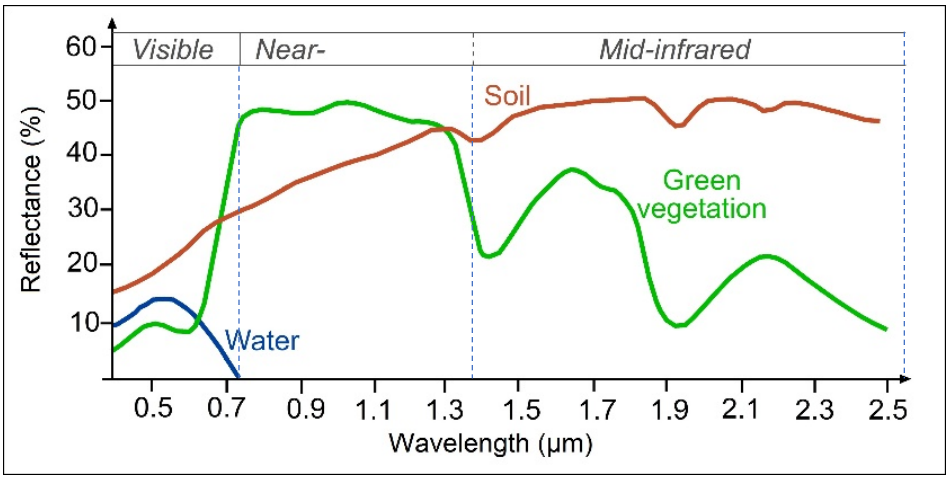



**Figure 2.5** Reflectance curve from multispectral image for identifying soil and green vegetation (modified after Huete, 2004).

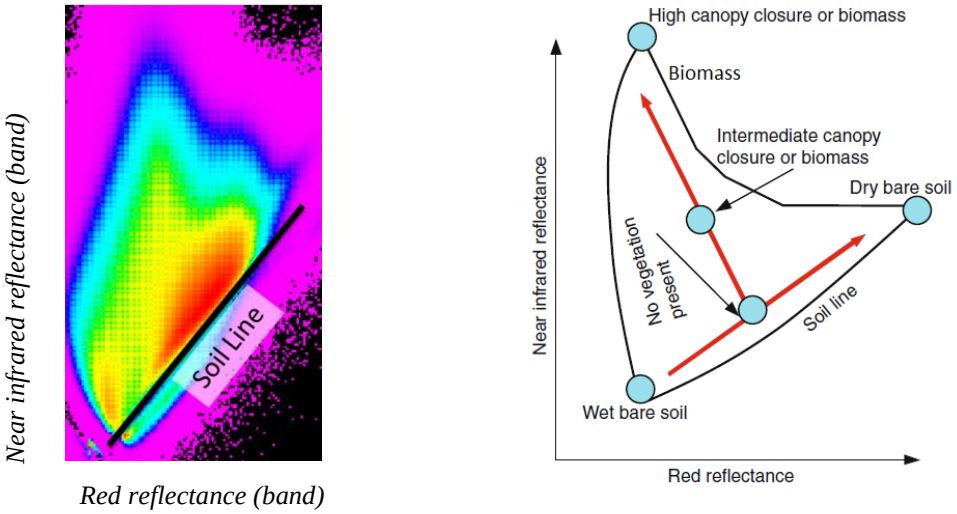



**Figure 2.6** Soil line indicated (in left and right side) from Red and Near infrared (NIR) band ratio (modified from Salama, 2011 and Qi *et al.*, 1994).

Most proximal crop sensors (*e.g.*, Chlorophyll meter, GreenSeeker, and Crop circle, etc.) produce local-scale data for experimental farms and have limited estimation capability for large scale precise nutrient deficiencies based on soil and crop reflectance (Ali *et al.*, 2015). These sensors rely on additional reference data to achieve their best results (Mulla, 2013). Recent research shows that low-altitude remote sensing methods are more successful at surface soil parameter prediction and crop stress approximation for large farms (Zhou *et al.*, 2016). Available

spectral bands from low-altitude sensing technologies have been used to simultaneously assess water stress and nitrogen requirements at different crop growth stages, as well as to evaluate soil characteristics, in a single field (Gregory *et al.*, 2006; Asher *et al.*, 2013). However, this needs to be validated across a diversity of fields.

A recent study by Wulf *et al.* (2015) demonstrated that the available visible multispectral bands and statistical approaches are unable to quantify soil minerals. To facilitate data analysis, several geospatial analysis tools and geostatistical approaches have been developed to leverage crop yields in large and local-scale precision farming (Cherlinka, 2017). Despite these initiatives, classification of multispectral remote sensing data has been proposed as a means to predict soil attributes and manage crop health. Unsupervised learning algorithms (fuzzy logic) applied to the multispectral data along with ground validation datasets have been used for bare soil mapping and management zone delineation (Cohen *et al.*, 2013). After reporting several limitations to the agricultural application of these methods, Belgui and Drăgu (2016) and Liu and Abd-elrahman (2018) used random forest supervised methods to report multisource data sensitivity, pattern recognition, and to classify thematic image maps for the prediction of soil classes.

Besides advanced methods to handle remote sensing data, data integration with other high-density PSS measurements has been proposed for precision farming (Hengl *et al.*, 2017; Albornoz *et al.*, 2018). Other efforts have integrated multispectral image indices and measured soil parameters which then could draw on historical yields for validation and estimation of crop biomass and potential crop yields (Gitelson *et al.*, 2003; Nguy-Robertson *et al.*, 2012). The most commonly encountered challenge in data fusion methods consists of the matching of spatial scale and accuracies at each level. Moreover, data processing and the cost of trading accuracy for better range in variability are also significant issues at the field-scale (Zhou *et al.*, 2016).

### 2.4 Use of sensor fusion in quantifying soil nutrients and solving agricultural issues

The application of individual sensor mapping and their analysis techniques in the context of agricultural soil mapping and crop management is limited because of the inability to measure a wide variety of sensor responses ranging from a soil's profile to its agronomic properties (Adamchuk and Viscarra Rossel, 2011). Real, or near-real, time sensor data (Huang *et al.*, 2018)

and the fusion of measurements with environmental variables entail a demanding approach to accurately map soil (Mahmood *et al.*, 2012; Samuel-Rosa *et al.*, 2015). Alchanatis and Cohen (2013) and Mulla (2013) stated that advances in sensor fusion algorithms and models are a growing concern and that only some have been applied successfully to predicting soil nutrients and fertility status. The key issues in data fusion and data analysis processes are the synchronization of different parameters at various scales and accounting for their multiscale uncertainties in a geographical space (Grunwald *et al.*, 2011; Aldabaa *et al.*, 2015).

Sensor fusion models often incorporate multivariable high-density data to solve agricultural problems. After a rigorous assessment of the data structure and preprocessing to remove potential outliers, hierarchical and geospatial models are deployed to predict soil properties and variable rate applications at various spatial and temporal scales (Kaye *et al.*, 2008; Grunwald 2009; Castrignanò *et al.*, 2017, McFadden *et al.*, 2017). Multivariate regression modeling (OLS and GLS) are widely used to evaluate relationships between variables (Hurvich and Tsai, 1989). Regression kriging has been applied to assess the relationship between predictor variables and soil properties from subsample datasets (Meirvenne and Cleemput, 2006). Hengel *et al.* (2004) presented a spatial prediction map at a regional scale by using regression kriging methods, while Xu *et al.*, (2018) employed this method along with remote sensing spectral indices to estimate total nitrogen in two different locations. Moreover, tree-based sensor-fusion algorithms are widely used in bioinformatics for precise prediction and faster decision making (Grunwald, 2006).

### *2.4.1 Digital soil modeling and thematic mapping at local-scale*

Supervised learning algorithms are another multivariate and high-density data analysis approach to generate faster decision-making processes. The algorithm evaluates prediction efficiency through noise and error modeling. Classification and regression tree (CART) modeling is a far more powerful method for spatial prediction and soil attribute mapping than simple, or multiple, linear regression (Bishop and Minasny, 2006). Besides various applications of CART data classification and regression methods in medical and remote sensing analyses, random decision forests and regression trees are drawing increasing attention in making thematic soil maps (Figure 2.7). Previous studies have indicated that this type of model deals effectively with unbalanced/missing datasets; it is more stable at a faster runtime and is robust for weighting

classified samples iteratively in remote sensing data analysis (Belgiu and Drăgu, 2016). Moreover, it allows for control of the variable selection from the training samples and shows an efficient error handling capability (Belgiu and Drăgu, 2016; Pelletier *et al.*, 2016). Hengl *et al.* (2004) applied such a model, along with sensor data fusion, to predict a wide range of soil-vegetation properties, as well as generating thematic maps on a regional scale. They analyzed various environmental covariates and then used input training samples for the machine learning techniques. In other studies, a spatial prediction framework was used to optimize the model parameters and reduce prediction uncertainties for predicting soil nutrients (Xiong *et al.*, 2015; Dharumarajan *et al.*, 2017; Merrill *et al.*, 2017; Vaysse and Lagacherie, 2017). While such methods have been adopted for regional scale prediction (Minasny and McBratney, 2016), there remains a need to implement a regression model for farm-scale applications.

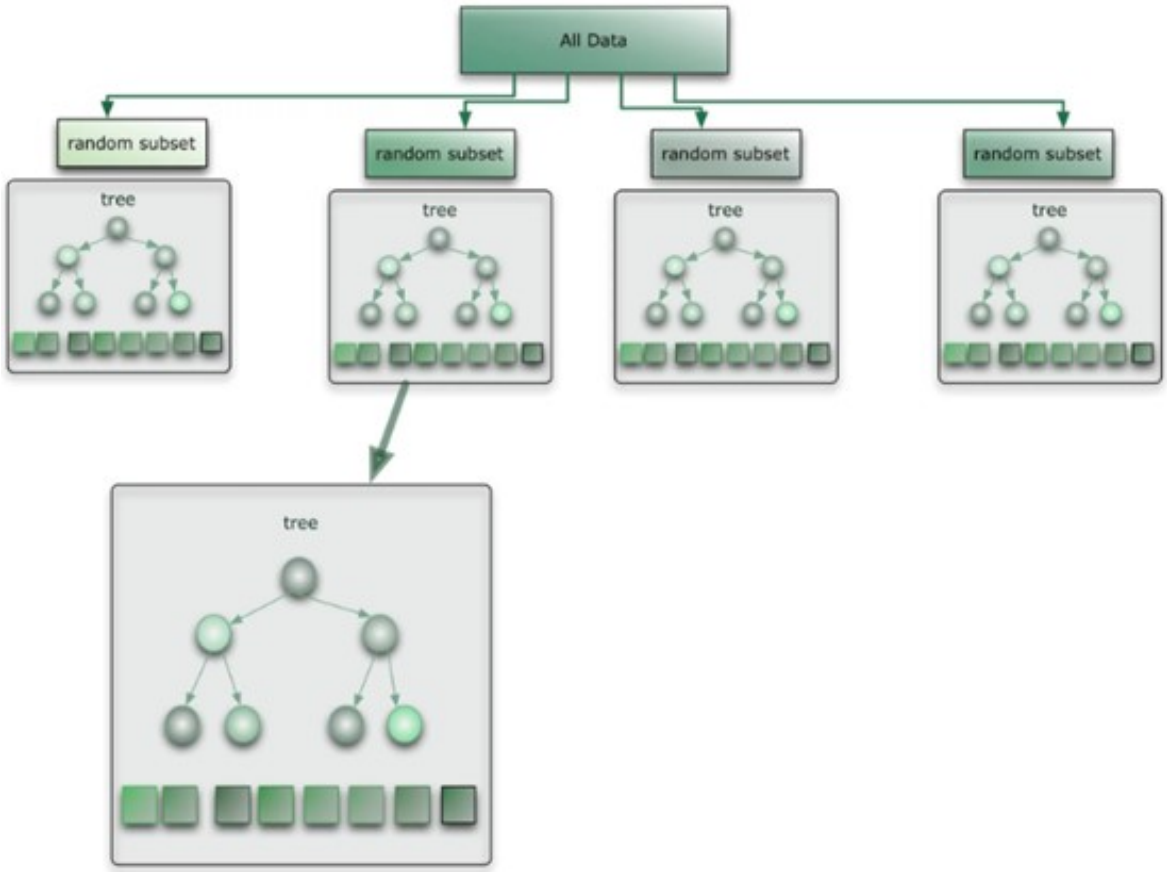



**Figure 2.7** Regression tree, an example of supervised decision tree model that optimizes the split from a small subset of training sets (Adopted from Géron, 2017).

The data fusion model offers the possibility of integrating geostatistical models to handle several environmental variables along with hierarchical relationships (Hengl *et al.*, 2004; Grunwald, 2009; Piikki *et al.*, 2013; Grunwald *et al.*, 2015). The random forest, a tree-model, has established the complex relationships among the variables and provided promising results in many ecological and environmental studies (Zhou *et al.*, 2019). The model's classification and regression approach can be applied in data mining and sensor data integration to solve agricultural problems, such as local-scale soil prediction and field-level accurate thematic maps. For this reason, model-based application rates are recommended to make faster decisions on accurate soil maps. In the decision-making steps, the models envision spatial variability and perform a complementary decision in maintaining soil management and its amendment requirements from historical farming practice. Moreover, faster decision-making enhances seasonal nitrogen management, amendments with organic matter, and management of topsoil for crop production (Grunwald *et al.*, 2011). Accurate model-estimated soil properties could help reduce agricultural inputs, making farms more profitable and sustainable by decreasing water and fertilizer consumption.

Modeling spatial and spatio-temporal data requires one to synchronize different parameters at various stages and handle their multiscale uncertainties in geographical space (Grunwald, 2009; Huang *et al.*, 2014). Geostatistical models, along with kriging/co-kriging, are useful in developing geospatial digital soil mapping (Hengl *et al.*, 2018); however, they require variogram parameters, anisotropic modelling parameters, fitting variograms using trends of covariates and link functions, etc. In contrast, the tree-based classification and regression models (*e.g.*, random forest) require limited user inputs to generate a thematic prediction map (Figure 2.8).

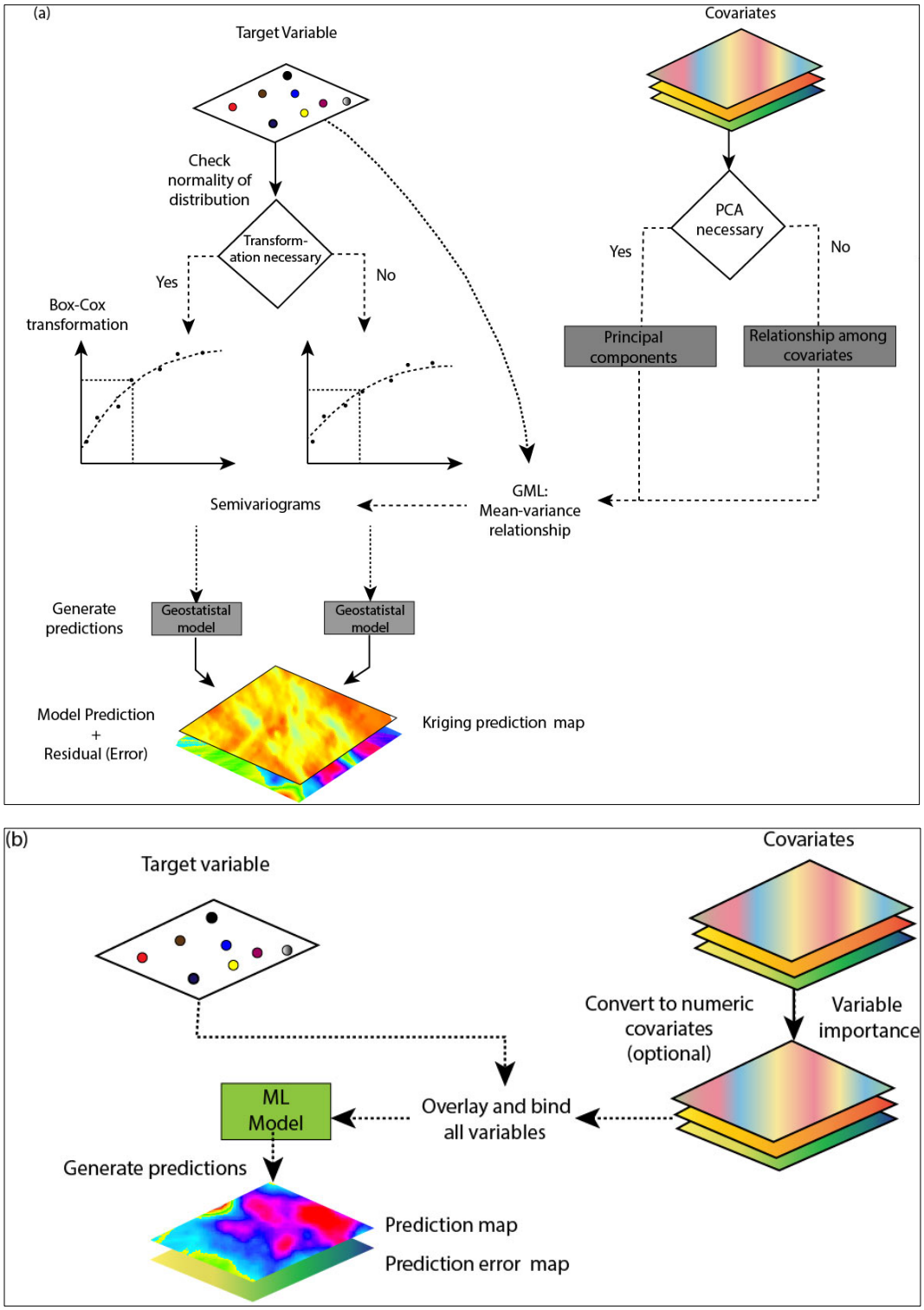



**Figure 2.8** Model-based geostatistics requires large amount of user inputs, such as specifies initial variogram parameters, anisotropy modeling, possibly transformation etc. (a), while classification and regression tree model requires only less user input (b) (Modified after Hengl *et al.*, 2018).

The above research depicted various challenges in implementing high-density sensor information and their efforts in thematic maps for precision agricultural practices. Most of the non-hierarchical clustering algorithms used in zone delineation only inadequately handle high-density data files with multiple variables and have limited accessibility. Farmers using precision farming practices often experience difficulties with variable rate operations due to fragmented management zones, in which those clustering techniques are often generated. In order to overcome the research gaps mentioned in section 2.1, the first study used high-density data to understand better field variability. Due to several limitations of the fuzzy clustering methods in the isolation process, the first objective of this study proposed a new multivariate clustering tool along with hierarchical methods to represent unique thematic maps and zonal boundaries based on the homogeneity of an agricultural field.

The PSS data, along with standard methods of laboratory soil analysis, are continuously assessing soil variability and plant available nutrient prediction in agricultural field management. The drawbacks discussed in Section 2.2, uncertainty analysis of predicting the soil properties is an emerging challenge in precision agricultural practice. To facilitate high-density sensor data application in farm management, uncertainty analysis of the prediction model and data quality are reported by different methods (RPD, SE, STD, $R^2$, etc.). In order to better assist in data quality assessment to delineate soil heterogeneity and their prediction capability, the second study of the thesis deals with the various aspects of proximal soil sensor (PSS) data through an assessment of the laboratory's data quality and comparing it with a wide range of lab-based measurements. This section also covers uncertainty analysis in model building and spatiotemporal soil mapping. As a result, lab testing results with a minimum of propagated errors are recommended for standard thematic soil mapping practices and for the laboratory proficiency certification program.

In precision agriculture, high-density proximal soil sensing (PSS) and remote sensing (RS) data are often used to predict thematic soil properties. Along with geospatial and multi-temporal sensor datasets, a subset of soil sampling data can also be drawn upon to predict soil nutrients in an agricultural field. Several limitations of multivariate data analysis techniques to the agricultural application were discussed in sections 2.3 and 2.4; the most commonly encountered challenges with data fusion methods consists of the matching of spatial scale and accuracies at each level. Regression tree-based sensor-fusion algorithms (i.e., random forest) require limited user inputs to

generate a thematic prediction map and are widely used in many environmental studies. The final study proposed the regression-based geospatial data integration model to establish the complex relationships among the different auxiliary variables along with the measured soil properties, and to delineate field variability and to assess prediction performance of different soil nutrients. Precise thematic maps can solve agricultural problems at a local scale.

**Connecting Text to Chapter 3**

The necessity for optimizing high-density data and field variability characterization of agricultural soils to support site-specific resource management was described in the previous chapter. Chapter 3 is related to the first objective as listed in Chapter 1 and the rationale illustrated in Chapter 2. In this chapter, an hierarchical data clustering technique was evaluated to determine the homogeneous parts of agricultural fields and to characterize field variability. The improvement of the effort was achieved by (1) implementing the uniform number of zones for optimizing soil sample locations and by (2) comparing with the results of traditional fuzzy clustering methods. To address the effectiveness of such an optimization, the proposed strategies were simulated using PSS-based dense apparent soil electrical conductivity ($EC_a$) data, topographic derivatives and RS-based vegetation indices. The proposed strategies were investigated in three agricultural fields.

Initial outcomes were reported and published at the conference proceedings and journal:

1. Saifuzzaman, M., & Adamchuk, V. (2017). Proximal Soil Sensing and Remote Sensing Data Processing for Precision Agriculture in Ontario, Canada. In *Abstracts from Annual Meeting of the Association of American Geographers, 5 - 9 April 2017* (pp. 1204–1205). Boston, Massachusetts, USA: (CD publication).
2. Saifuzzaman, M., & Adamchuk, V. (2017). Geospatial Analysis of Proximal Soil Sensing and Remote Sensing Data in Precision Agriculture. In *Abstracts from the Earth Observation Summit 2017, UQAM (Science Centre), 20 - 22 June 2017. Canadian Remote Sensing Society.* Montreal, Quebec, Canada: (Published on-line at https://crss-sct.ca/conferences/csrs2017).
3. Saifuzzaman, M., Adamchuk, V., Huang, H., Ji, W., Rabe, N., & Biswas, A. (2018). Data Clustering Tools for Understanding Spatial Heterogeneity in Crop Production by Integrating Proximal Soil Sensing and Remote Sensing Data. In *Proceedings of the 14th International Conference on Precision Agriculture, 24 - 27 June 2018. International Society of Precision Agriculture* (p. 14). Montreal, Quebec, Canada: (Published on-line at http://www.ispag.org).
4. Saifuzzaman, M., Adamchuk, V., Buelvas, R., Biswas, A., Prasher, S., Rabe, N., Aspinall, D., & Ji, W. (2019). Clustering Tools for Integration of Satellite Remote Sensing Imagery and Proximal Soil Sensing Data. *Remote Sensing – MDPI* 11(9).

# Chapter 3: Clustering Tools for Integration of Satellite Remote Sensing Imagery and Proximal Soil Sensing Data

Md Saifuzzaman, Viacheslav Adamchuk, Roberto Buelvas, Asim Biswas, Shiv Prasher, Nicole Rabe, Doug Aspinall and Wenjun Ji

**Abstract**

Remote sensing (RS) and proximal soil sensing (PSS) technologies offer an array of advanced methods for obtaining information on soil properties and for determining soil variability for precision agriculture. The large amount of data collected by these sensors may provide essential information for precision, or site-specific, management in a production field. Data clustering techniques are crucial for data mining, and high-density data analysis is important for field management. A new clustering technique was introduced and compared with existing clustering tools to determine the relatively homogeneous parts of agricultural fields. A DUALEM-21S sensor, along with high-accuracy topography data, was used to characterize soil variability in three agricultural fields situated in Ontario, Canada. Sentinel-2 data assisted in quantifying bare soil and vegetation indices (VIs). The custom Neighborhood Search Analyst (NSA) data clustering tool was implemented using Python scripts. In this algorithm, part of the variance of each data layer is accounted for by subdividing the field into smaller, relatively homogeneous, areas. The algorithm's attributes were illustrated using field elevation, shallow and deep apparent electrical conductivity ($EC_a$), and several VIs. The $R^2$ of the k-means cluster relative to that of the NSA was higher in most of the fields; it was approximately 0.80. The k-means cluster map consisted of pixels with isolated boundaries in various parts of a field, whereas the NSA algorithm reduced zone fragmentation and produced spatially congruous zones. The unique feature of this proposed protocol was the successful development of user-friendly and open source options for defining the spatial continuity of each group and for use in the zone delineation process.

**Keywords:** remote sensing; proximal soil sensing; clustering techniques; spatial homogeneity; management zones.

## 3.1 Introduction

A delineated areal extent (DAE) is a finite part of a field representing a unique and homogeneous portion of data [1–2]. The determination of DAEs, or zones, using remote sensing (RS) and proximal soil sensing (PSS) data is becoming critical in the assessment of soil properties and the characterization of variability in precision agriculture [1–8]. In the delineation process, high-resolution data from these sensing technologies, together with quantitative methods, are used to infer the spatial pattern of soil heterogeneity [9–13]. To obtain information on the spatial pattern of soil parameters and produce thematic soil maps to understand a field's agronomic and yield-limiting factors, high-density and multivariate data analyses were drawn upon to isolate homogeneous field areas and to identify potential management zones [14–20].

Multivariate data and hierarchical clustering techniques are crucial for identifying and understanding soil variability within a production field [13, 21–25]. Among the multivariate data analysis techniques, the unsupervised clustering techniques of fuzzy c-means and k-means are most commonly used for data mining [26–32]. Because of the fuzziness of c-means and k-means and other limitations in the isolation process, each cluster object can belong in more than one group and boundary pixels are created [8,33,34]. This study attempted to provide a multivariate and hierarchical clustering tool to represent unique thematic maps, and zonal boundaries based on the homogeneity of the agricultural field.

Most clustering algorithms applied in zone delineation do not handle high-density data files with multiple variables [35–39] nor do they produce an optimal number of zones. As clustering techniques commonly generate fragmented management zones, agricultural scientists and farmers face challenges when implementing variable-rate operations [8, 16, 40–44]. In practice for field operations, the optimal number of zones should be such that the capacity of GPS-guided field equipment is neither overtaxed (too many isolated zones) nor underexploited (too few isolated zones). A survey conducted using a Real-Time Kinematic (RTK), DUALEM proximal soil sensor, and a remote sensing satellite sensor yielded high-density elevation, apparent electrical conductivity ($EC_a$), and surface vegetation reflectance data, respectively. In this research, the proposed data clustering algorithm was optimized to generate spatially contiguous zones to aid in the achievement of best management practice goals. This study presents the process used to develop a new and enhanced clustering technique to better understand soil variability (e.g.,

topography, crop performance, and high-density soil data, such as $EC_a$), in an agricultural field. The performance of this technique was then compared to that of other commonly used techniques.

## 3.2 Materials and Methods

### *3.2.1 Experimental Sites and Data Description*

Situated at the Woodrill Farms near Guelph, Ontario, Canada, three agricultural fields (namely, WH, LD, and RB), differing in acreage and soil class, were surveyed using both RS and PSS sensors (Table 3.1 and Figure 3.1). The PSS equipment was pulled behind an all-terrain vehicle; it measured elevation and $EC_a$ data points (collected between August 2015 and April 2016) for the experimental sites at intra- and inter-row spacing of 5 m and 10 m, respectively. Elevation data points were collected by an RTK Global Navigation Satellite Systems (GNSS) receiver (Trimble Inc., California, USA) (Table 3.2). On the basis of the high-density elevation points, a digital elevation model (DEM) was created with a spatial precision of about 2 cm horizontally and 3 cm vertically. Slope, aspect ratio [sin(aspect/2)], and a topographic wetness index (TWI) were derived from a DEM of the study sites. Developed by Beven and Kirkby [45] and serving to investigate hydrological processes controlled by topography, the TWI was determined using the SAGA GIS v.2.4 (University of Hamburg, Germany). TWI = $[\ln \frac{a}{\tan \beta}]$ where; $a$ is the upland contributing area, $[(\text{flow accumulation} + 1) \times \text{cell size}]$, and $\beta$ is the slope in radians.

The DUALEM-21S system (DUALEM Inc, Milton, ON, Canada) had one transmitter coil and four receivers—two of horizontal coplanar (HCP) geometry and two of perpendicular coplanar (PRP) geometry—at a separation distance of 1 to 2 m. It was used to collect $EC_a$ at four different depths: PRP1 at 0–0.5 m, PRP2 at 0–1.0 m, HCP1 at 0–1.6 m, and HCP2 at 0–3.0 m (Table 3.3). The pre-processing procedures for the collection of RTK elevations and $EC_a$ values were similar and included raw data display, the identification of missing values, median filtering, and the removal of outliers. Culled data included: (i) start pass and end pass delays, (ii) points with over speed limits, (iii) values outside the user-defined minimum and maximum values and (iv) changes in pitch or roll outside the acceptable limit. Data outliers were removed on the basis of the above criteria, such that about 15% of data points were removed. Various methods of geospatial data processing were undertaken on multiple data layers, including rectification, interpolation, and point data extraction. These methods enhanced the data quality for further analysis.

**Table 3.1** Characteristics of three agricultural fields in Guelph, Ontario, Canada.

| Field ID | Area (ha) | Soil classes | Target crops |
|---|---|---|---|
| WH | 39.60 | Loam | Soybean/Wheat |
| LD | 21.00 | Sandy Loam | Soybean |
| RB | 75.00 | Fine Sandy Loam | Soybean/Wheat |

**Table 3.2** Summary statistics of elevation data from the Real-Time Kinematic (RTK) sensor for three agricultural fields in Guelph, Ontario, Canada. Number of sensor measurements varied based on the experimental sites and sensor settings (data points recorded every 0.1s and in parallel lines of about 12m separation).

| Field ID | # of measurements | Elevation (m) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Min | Median | Max | Range | STD | Mean |
| WH | 28493 | 372.06 | 378.07 | 384.54 | 12.48 | 2.33 | 378.21 |
| LD | 7110 | 332.70 | 344.86 | 354.17 | 21.47 | 5.76 | 343.95 |
| RB | 20813 | 358.41 | 367.67 | 372.16 | 13.75 | 3.63 | 366.64 |

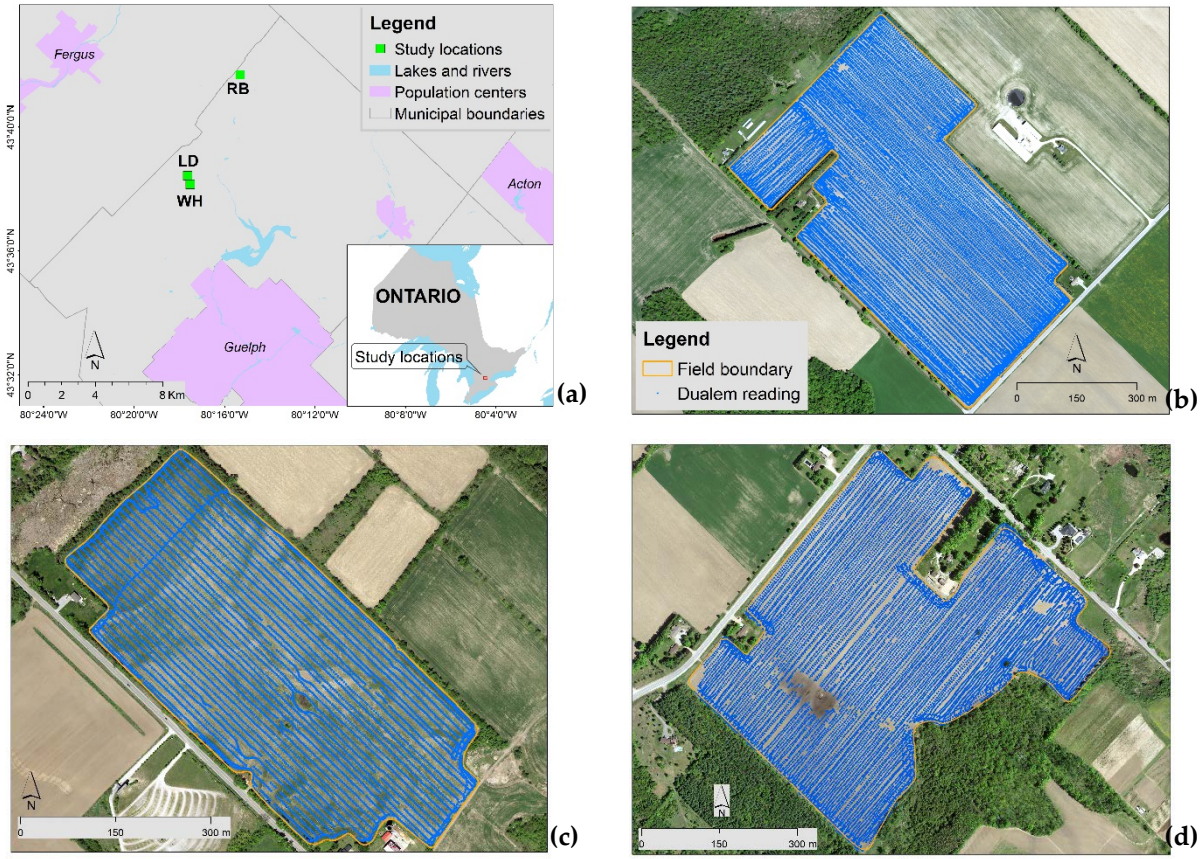



**Figure 3.1** (**a**) Location and aerial views of three fields at the Woodrill Farms in Guelph Ontario, Canada: WH field boundary with soil apparent electrical conductivity ($EC_a$) data points (**b**), LD field boundary with soil $EC_a$ data points (**c**), and RB field boundary with soil $EC_a$ data points (**d**).

**Table 3.3** Summary of statistics from DUALEM-21S sensor readings from the three agricultural fields. HCP: horizontal coplanar, PRP: perpendicular coplanar. Number of sensor measurements varied based on area of experimental sites and sensor settings at the data collection time (data points recorded every 0.1s and in parallel lines of about 12m separation).

| Field ID | # of Measurements | Sensor Configuration | Apparent Soil Electrical Conductivity ($EC_a$), mS $m^{-1}$ | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Min | Median | Max | Range | STD | Mean |
| WH | 20129 | | 4.00 | 12.28 | 25.28 | 21.28 | 1.69 | 12.51 |
| LD | 6931 | HCP1 | 2.58 | 6.90 | 16.08 | 13.50 | 1.55 | 6.96 |
| RB | 18524 | | 1.70 | 9.00 | 17.98 | 16.28 | 2.81 | 9.13 |
| WH | 20129 | | 4.68 | 7.92 | 22.24 | 17.56 | 1.60 | 8.15 |
| LD | 6931 | PRP1 | 0.72 | 4.44 | 14.12 | 13.40 | 1.38 | 4.55 |
| RB | 18524 | | 0.00 | 3.53 | 16.80 | 16.80 | 2.86 | 4.40 |
| WH | 20129 | | 7.42 | 10.46 | 24.42 | 17.00 | 1.79 | 10.83 |
| LD | 6931 | HCP2 | 0.50 | 4.44 | 14.44 | 13.94 | 1.85 | 4.61 |
| RB | 18524 | | 2.50 | 8.45 | 14.99 | 12.49 | 2.65 | 8.22 |
| WH | 20129 | | 5.42 | 9.10 | 23.92 | 18.50 | 1.75 | 9.37 |
| LD | 6931 | PRP2 | 1.08 | 4.68 | 14.60 | 13.52 | 1.50 | 4.75 |
| RB | 18524 | | 0.14 | 5.10 | 15.00 | 14.86 | 2.96 | 5.64 |

A Sentinel-2 image was used to analyze bare soil and vegetation characteristics (Table 3.4). Remote sensing image processing steps were followed, including radiometric correction, stitching, co-registration, and stack bands. One OrthoPhoto and two Sentinel-2 images were used for co-registration and visual interpretation with zonal thematic maps. In addition to the traditional visible (RGB) and near-infrared (NIR) spectral bands, Sentinel-2 imagery presented three red edge parts of the spectrum as well, where only the red-edge B5 (704 nm) band was used for further analysis. Spectral indices were produced from Sentinel-2 data to identify the strong absorption spectrum of chlorophyll. These included the Difference Vegetation Index (DVI), the Normalized Difference Red Edge Index (NDRE), the Normalized Difference Vegetation Index (NDVI), and the Modified Soil Adjusted Vegetation Index (MSAVI2). Among the vegetation indices (VIs), NDVI maps were found to be more suitable at the early crop growth stage and were used for the clustering process [46, 47].

**Table 3.4** Remote sensing data characteristics and their sources.

| Satellite Sensor | Spectral Bands | Pixel (m) | Central Wavelength(nm) | Imaging Date | Source |
|---|---|---|---|---|---|
| OrthoPhoto | B, G, R, NIR | 0.2 | - | 23 May 2015 | OMAFRA/OMNRF[1] |
| Sentinel-2 | 2(B), 3(G), 4(R), 8(NIR) | 10.0 | 494, 560, 665, 834 | 21 July 2017 | Planet Labs |
| Sentinel-2 | 5,6,7 (red edge 1,2 &3) | 20.0 | 704, 740, 781 | 21 July 2017 | Planet Labs |

[1]Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) and Ontario Ministry of Natural Resources and Forestry (OMNRF).

### *3.2.2 Interpolated Maps of Selected Sensor Variables*

Ordinary Kriging interpolation maps were generated from the PSS measurements in ESRI ArcGIS software (v10.5.1). Kriged maps (with a spatial resolution of 5 m) showing RTK elevation (DEM), derived topographic variables (including slope, aspect, and TWI), and DUALEM sensor variables (HCP1, HCP2, PRP1, and PRP2) were produced. Slope and aspect showed similar field patterns as TWI and thus, were deemed redundant. In the final clustering process only TWI was used. Due to fewer saturation problems at early crop growth stage in the fields and similar results in NDRE, widely used NDVI maps (with a spatial resolution of 10 m) were extracted for the clustering tool. Those maps represented significant variations across the expanse of each field (Figures 3.2, 3.3, and 3.4). The interpolated maps were extracted into a data file of multiple layers. Finally, a text data file was generated to store the sensor-derived variables for input into the newly developed clustering tool and commonly used fuzzy clustering techniques.

To delineate zones, the multilayer data files were analyzed by the proposed data clustering tool. The new data clustering algorithm and its processing steps are elaborated in detail in the following section, as well as the new algorithm's clustering outputs in comparison to outputs from fuzzy clustering techniques.

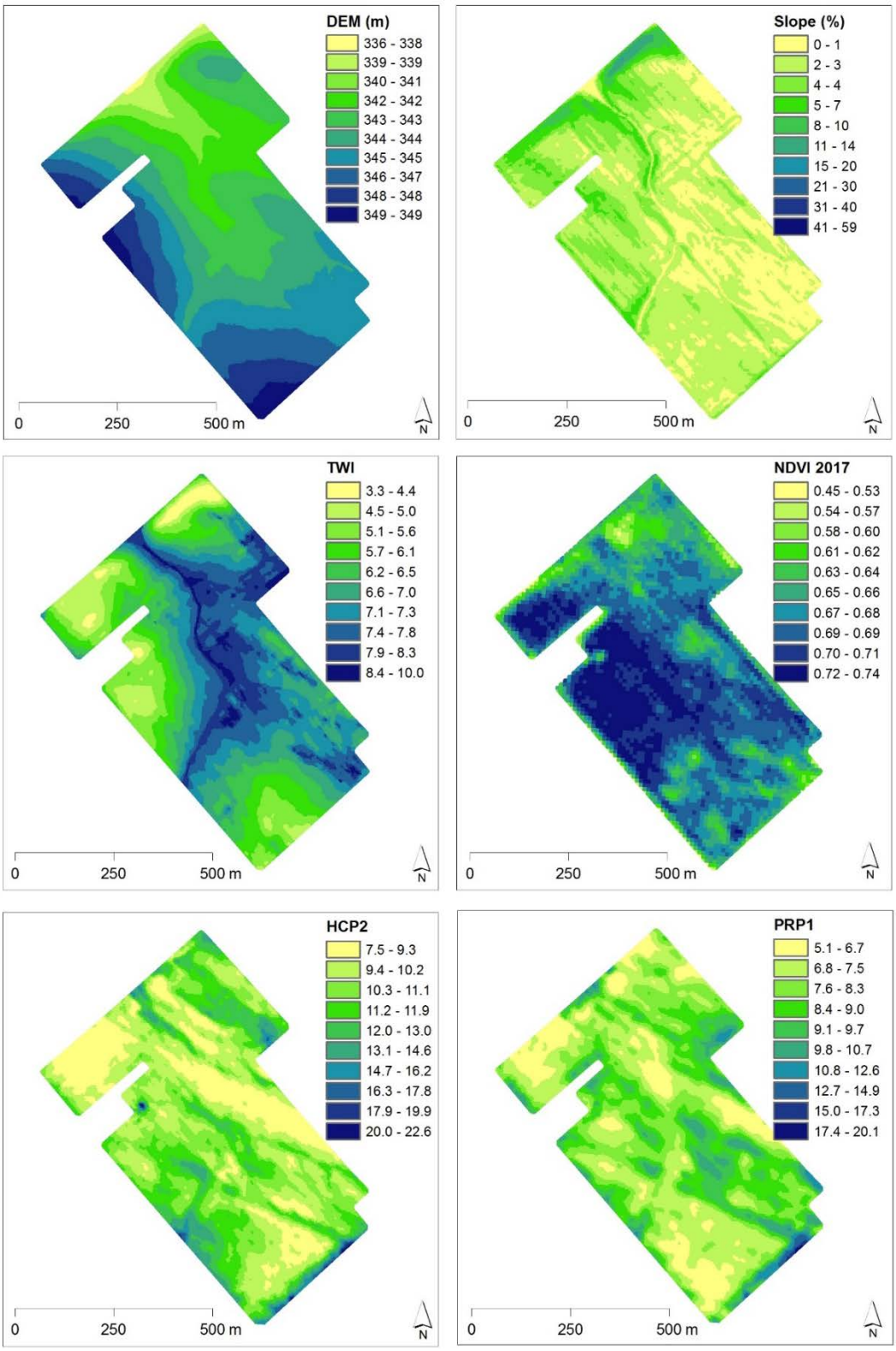



**Figure 3.2** Interpolated maps (Kriged) of digital elevation model (DEM), topographic wetness index (TWI), two apparent electrical conductivity measurements (HCP2 and PRP1), and Normalized Difference Vegetation Index (NDVI) maps for the WH field.

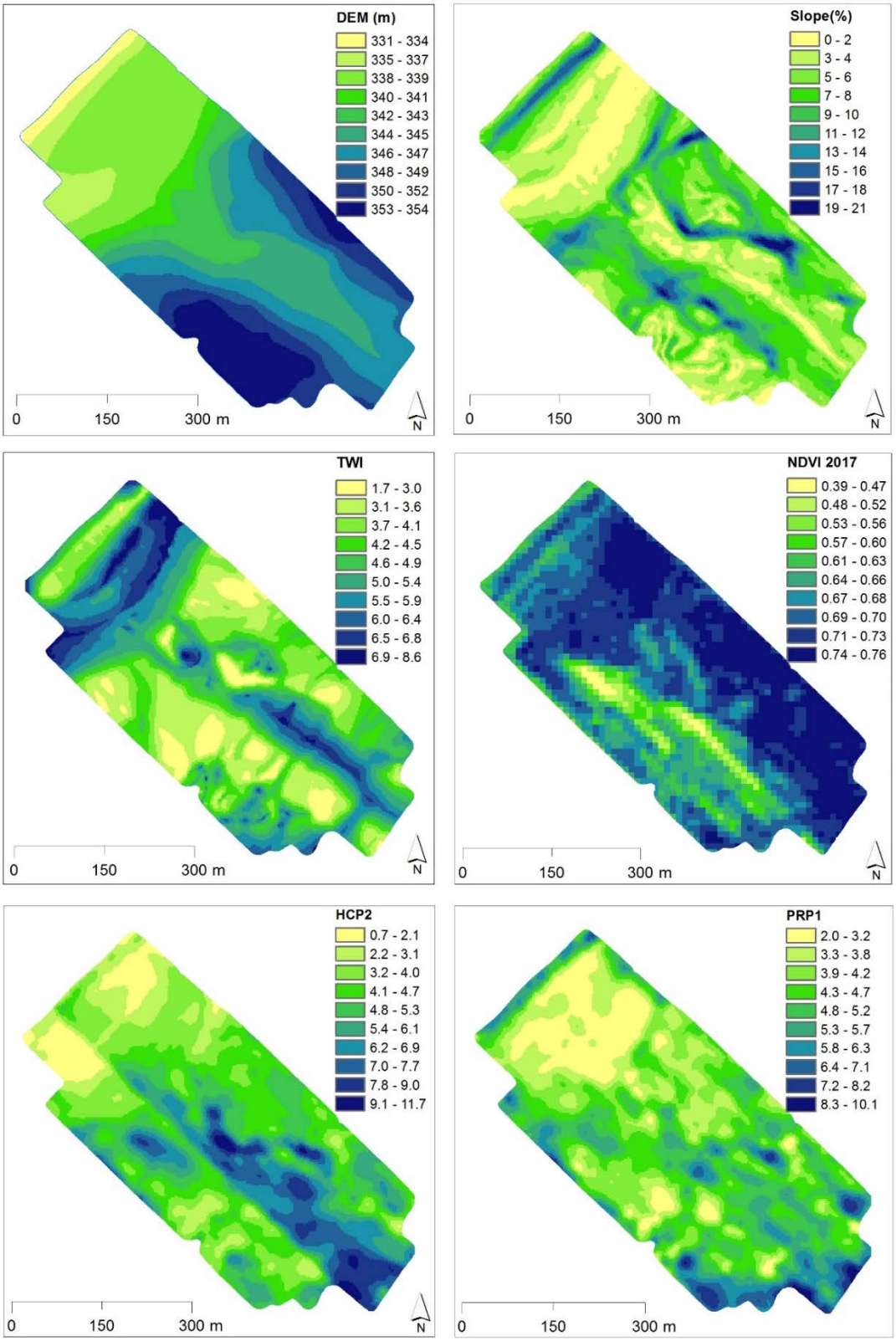



**Figure 3.3** Interpolated maps (Kriged) of digital elevation model (DEM), topographic wetness index (TWI), two apparent electrical conductivity measurements (HCP2 and PRP1), and Normalized Difference Vegetation Index (NDVI) maps for the LD field.

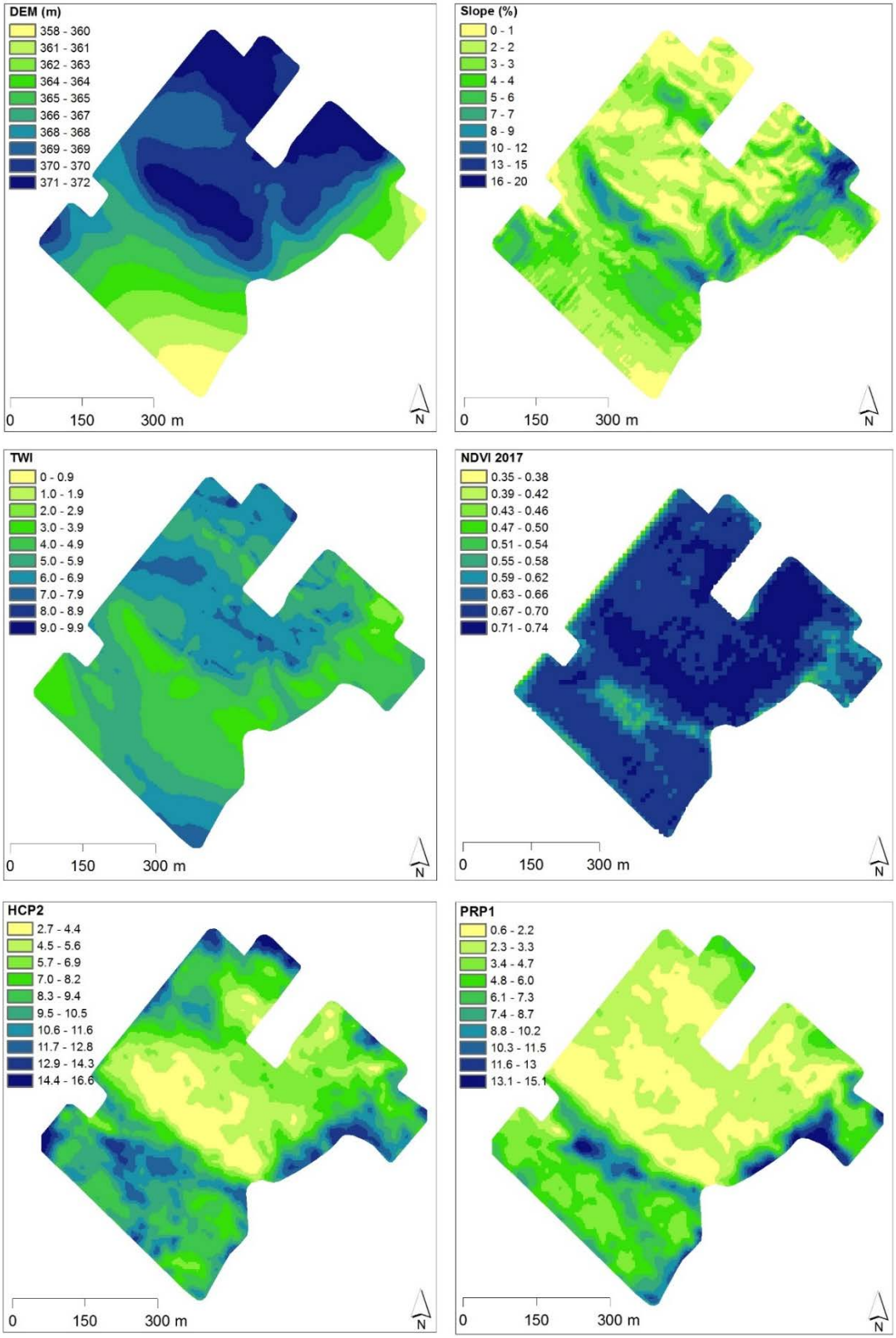



**Figure 3.4** Interpolated maps (Kriged) of digital elevation model (DEM), topographic wetness index (TWI), two apparent electrical conductivity measurements (HCP2 and PRP1), and Normalized Difference Vegetation Index (NDVI) maps for the RB field.

*3.2.3 Data Clustering Algorithms*

Fuzzy c-means calculated by the management zone analyst (MZA) [48] were used to generate the normalized classification entropy (NCE) and fuzziness performance index (FPI) of the five zones. Due to the limitations of handling several multiple layers for creating a large number of zones, MZA produced only five zones in this study. The k-means algorithm in the Python data library was used to generate (k = 5, k =15, and k = 25) clusters and cluster centers were determined using the sum of square distances of all data points and the number of cases in each cluster. Initially, five user-defined clusters were defined in the above clustering methods; however, the optimum number of zones was determined in the final step and compared between the two methods.

The proposed data clustering method, called the Neighborhood Search Analyst (NSA), resulted in the algorithms shown in Figure 3.5. The processing steps and formula were adopted from the NSA and were written in MATLAB scripts [6]. Preliminary tests of the algorithm in numerous production fields highlighted the algorithm's robustness when partitioning field areas using several field measurements. To construct an objective function to be optimized through the data grouping process, the mean squared error (MSE) was calculated for each individual data layer k according to:

$$MSE_k = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_j\right)^2}{N - m} \tag{1}$$

where $X_{ij}$ is a sensor value for the $i_{th}$ grid cells within the $j_{th}$ group; $\overline{X}_j$ is the mean of $j_{\text{th}}$ group; $N$ is the total number of grid cells; $m$ is the number of groups; and $n_j$ is the number of grid cells within the $j_{th}$ group.

It should be noted that the difference between the total number of grid cells and the number of groups can be determined by:

$$N - m = \sum_{j=1}^{m}\left(n_j - 1\right) \tag{2}$$

Since the algorithm initially assumes that all grid cells belong to the same group, labeled "1" and designated as "the rest of the field", then $MSE_k(m=1)$ represents the variance of the $k_{th}$

data layer across the entire field. Given that the area of the field is substantially greater than the area of a grid cell, $MSE_k(m=1)$ can be termed Farthest Distance Variance ($FDV_k$). In such a situation, the portion of data variance accounted for by distributing $N$ grid cells among $m$ groups can be calculated as:

$$R_k^2 = 1 - \frac{MSE_k}{FDV_k} \tag{3}$$

where $MSE_k(m=1)$ can be called Farthest Distance Variance ($FDV_k$).

The maximum value of $R^2_k$ can be obtained when $MSE_k$ is as small as possible. It approaches 1 when the number of groups increases. Since the result can be considered less favorable if at least one data layer $k$ is not adequately accounted for, it is reasonable to employ the integration operator OR instead of the more common AND. This avoids the need to assign a weight factor to each individual data layer when adding corresponding $MSE_k$ estimates. In mathematical terms, this would mean that the product of all $R^2_k$ should be maximized. Therefore, the objective function (OF) was defined as:

$$OF = \prod_{k=1}^{K} R_k^2 \tag{4}$$

where $K$ is the number of PSS data layers.

In this study, the smallest number of data elements that could be grouped within the grid cell square window was nine (3×3). Therefore, the maximum accountable variance is the variance of PSS measurements between immediate neighbors. The Shortest Distance Variances ($SDV_k$) can be found using:

$$SDV_k = \frac{1}{w} \sum_{j=1}^{w} \sum_{i=1}^{9} \frac{\left(X_{ij} - \overline{X}_j\right)^2}{8} \tag{5}$$

where $w$ is the total number of 3x3 square windows of grid cells.

Since $SDV_k$ represents the smallest $MSE_k$ value, the maximum value of $R^2_k$ is calculated as:

$$R^2_{k\,\max} = 1 - \frac{SDV_k}{FDV_k} \tag{6}$$

This $R^2_{k\,max}$ parameter can range between 0 and 1. It is equal to 0 when data layer $k$ is either uniform or highly variable, so that $SDV_k = FDV_k$. In such a case, the data layer should not be able

to affect changes in the OF. Alternatively, when $R^2_{k\,max}$ is close to 1, the data layer has a strong spatial structure ($SDV_k << FDV_k$), and OF must be sensitive to the change of $MSE_k$ corresponding to that particular data layer. In mathematical terms, this goal can be achieved by multiplying all $R^2_k$ values raised to the $R^2_{k\,max}$ power of:

$$OF = \prod_{k=1}^{K} R_k^{2\,R^2_{k\,max}} = \prod_{k=1}^{K} \left(1 - \frac{MSE_k}{FDV_k}\right)^{\left(1-\frac{SDV_k}{FDV_k}\right)} \tag{7}$$

The resultant OF indicates the overall quality of grid cell groupings. It varies from 0 to 1 and approaches high values when every spatially structured layer of the PSS measurements is separated among spatially continuous groups of grid cells with minimum internal group variance. Such groups represent different combinations of average PSS measurements obtained with different sensors that diverge from average field conditions. To facilitate the formation of grid cell groups that would maximize the OF, the NSA algorithm was implemented in this study using Python v3.6 (created by Guido van Rossum and managed by Python Software Foundation, Delaware, USA).

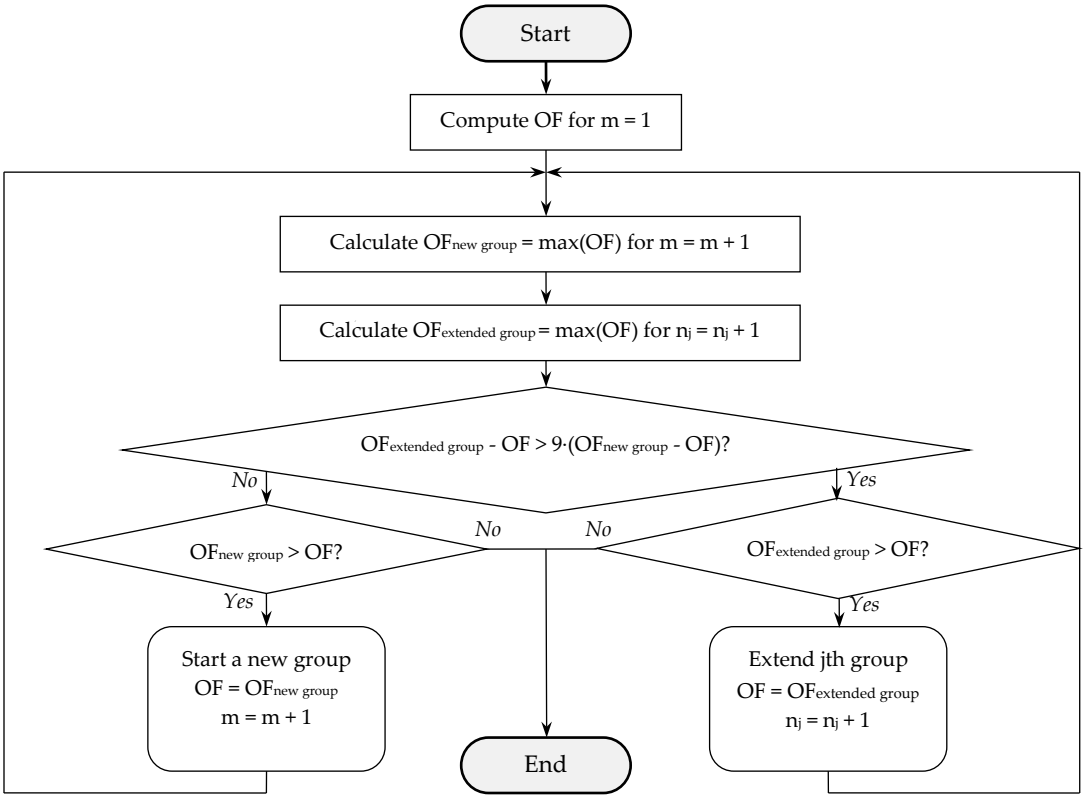



**Figure 3.5** The flowchart of the Neighborhood Search Analyst (NSA) algorithm process.

### 3.3 Results and Discussion

*3.3.1 c-Means Clustering*

On the basis of the seven input variables (*i.e.*, elevation, TWI, NDVI, HCP1, HCP2, PRP1, and PRP2) of the WH field, Euclidean distance-based NCE and FPI indices in FCM clustering were assessed for their performance in creating an optimum number of zones. NCE and FPI reached the maximum values in either 4 or 5 zones (Figure 3.6). This clustering method is flawed when it comes to obtaining an optimum number of zones [8, 49, 50]. The FCM clusters produced pixels with isolated boundaries in various parts of the field [51, 52]. Many studies have reported this representation problem regarding the clustering of data due to the fuzzy boundary [16, 32, 53, 54]. In the present method, user-defined numbers of clusters were produced without considering the geospatial locations of the dataset (spatial continuity) or their distances.

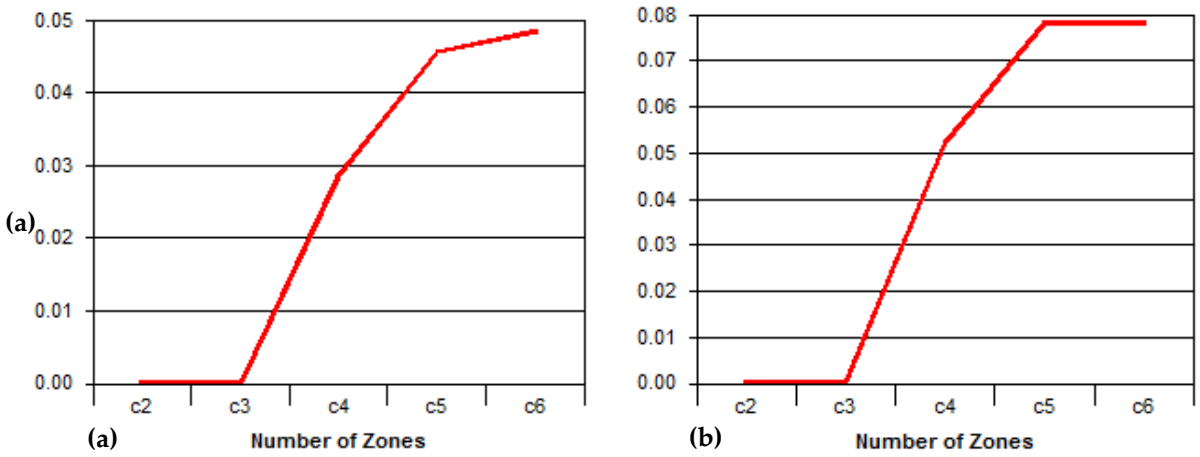



**Figure 3.6** Normalized classification entropy **(**NCE) (**a**) and fuzziness performance index (FPI) (**b**) of the WH field based on seven input variables.

*3.3.2 k-Means Clustering*

In the k-means clustering (k=5), the data values were taken directly from the input table of the WH field for generating cluster centers (Figure 3.7a). Data were standardized and normalized for the specific variable values. Among the five user-defined clusters, clusters 1, 2, 3, and 5 used the most data points. The variation among the zones was understood where maximum data points/variable values used to yield. Since there was a random component, after several runs of each clustering process, the coefficient of determination ($R^2$) varied according to how the k-means

algorithm was initialized. The cluster map consisted of groups of pixels with isolated boundaries in various parts of the WH field (Figure 3.7b). Figure 3.7b shows that the k-means cluster map of the WH field generated 36 scattered zones of user-define clusters (k=25).

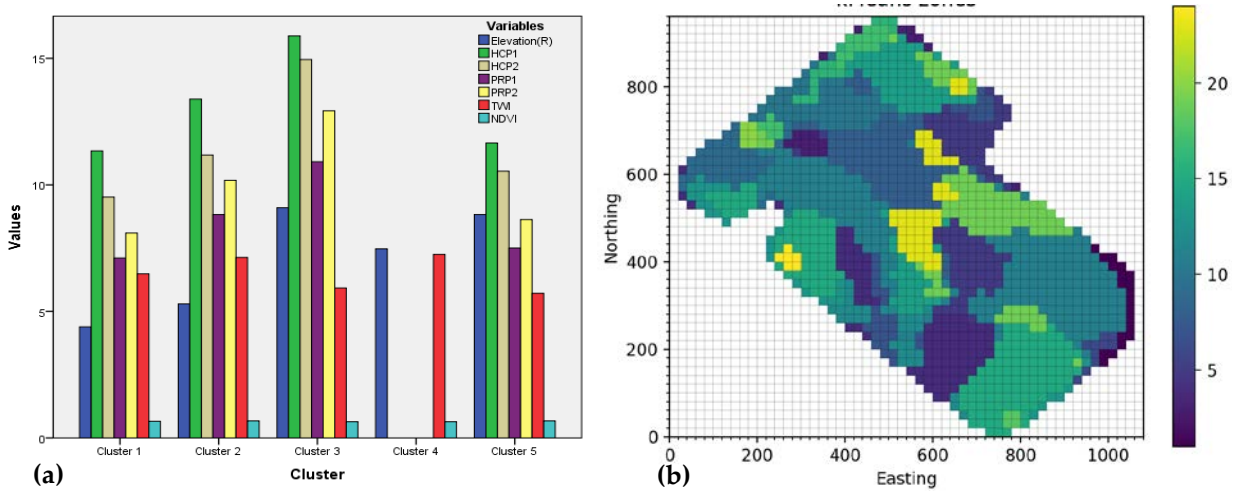



**Figure 3.7** (**a**) k-means cluster (k = 5) centers with variable values of the WH field and (**b**) k-means cluster (k = 25) map of the WH field showing zones with various isolated pixels.

### *3.3.3 NSA Clustering*

In the NSA zone delineation process, unlike other clustering algorithms, providing the number of field partitioning clusters is not obligatory. Without defining the number of clusters, NSA produced an optimum number of groups for the grid cell (grid size of 20 m), separately, for seven different input variables. According to the efficiency of the clustering algorithms, the best possible, or maximum number of variations (optimum zones) were detected in the fields. More importantly, this clustering tool efficiently delimited maps by providing the optimum number of zones for field management (Figures 3.8a, 3.9a, and 3.10a). On this basis, the WH, LD, and RB fields have 28, 20, and 27 georeferenced zones, respectively. For NSA clustering, user-defined (k = 5, k = 15, and k = 25) zones were delineated and are illustrated later in this paper.

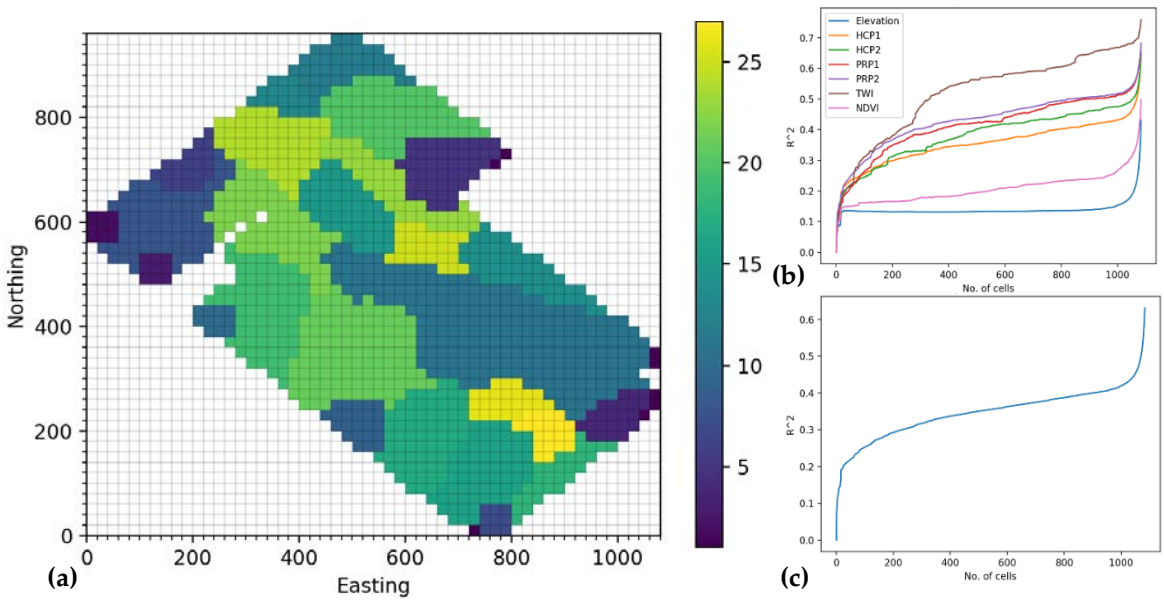



**Figure 3.8 (a)** Zonal map including 28 well-defined clusters; (**b**) Coefficient of determination ($R^2$) for each data layer; and (**c**) Overall objective function (OF) vs number of grid cells (WH).

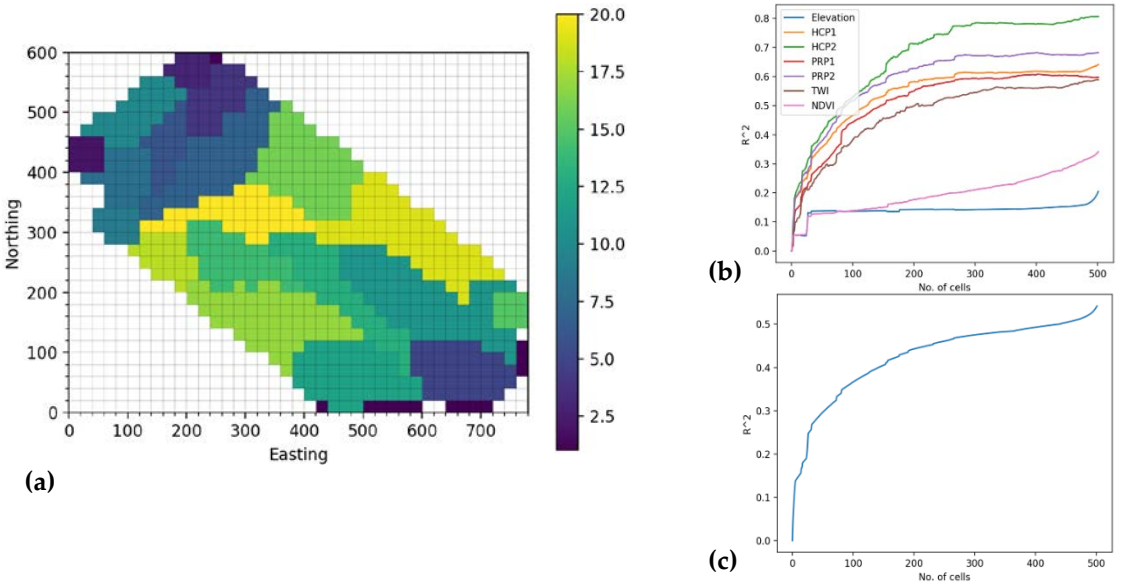



**Figure 3.9** (**a**) Zonal map including 20 well-defined clusters; (**b**) Coefficient of determination ($R^2$) for each data layer; and (**c**) Overall OF vs number of grid cells (LD).

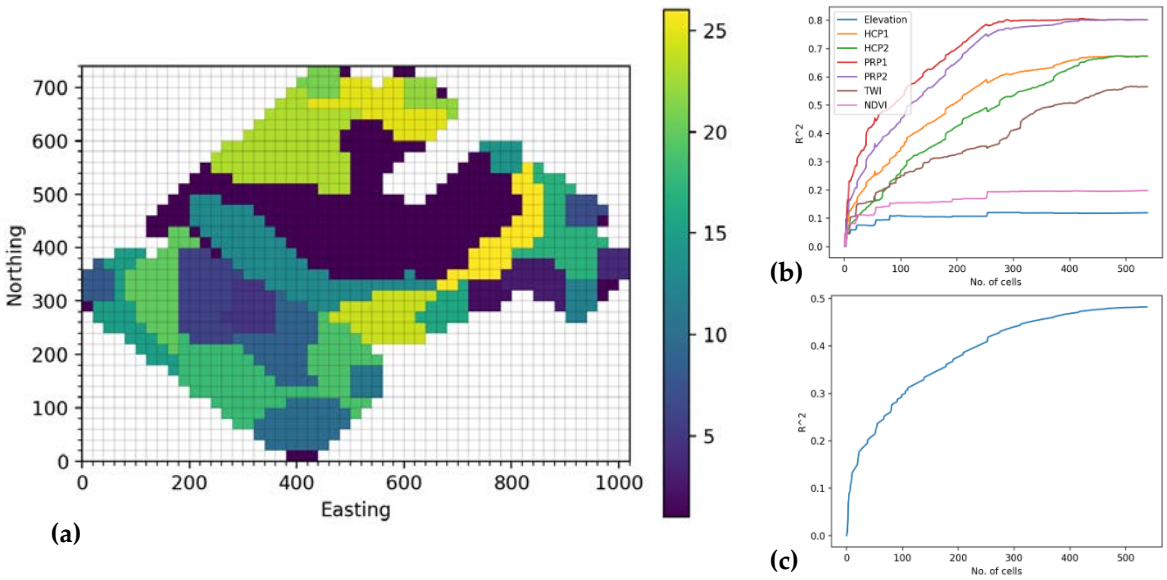



**Figure 3.10** (**a**) Zonal map including 27 well-defined clusters; (**b**) Coefficient of determination ($R^2$) for each data layer; and (**c**) Overall OF vs number of grid cells (RB).

In NSA, zone delineation was performed by the individual $R^2$ values of each variable (Figures 3.8b, 3.9b, and 3.10b) and overall OF (Figures 3.8c, 3.9c, and 3.10c). These graphs show the part of the variance of each data layer which was accounted for by subdividing the field into smaller areas. The software also decided which variables (among the seven input variables) contributed more variations and used them for making a homogeneous number of zones. In this study, NDVI and elevation parameters had a low contribution in creating the zones. In each graph, the greater $R^2$ value indicated that variability within individual zones was smaller than the difference between zones. Figures 3.8b, 3.9b, and 3.10b show that the $R^2$ values increased when new groups were formed or added to the existing groups. The NSA that produced $R^2_{max}$ value was about 0.9, and the graph had a steeper initial slope. This indicated that the data layer had a strong spatial structure and was dominant when the field was split. Moreover, the x value (No. of cells), where most graphs leveled off, showed that the smallest level of field partitioning revealed most of the soil heterogeneity. Results in LD and RB fields indicated that $R^2$ for each data layer reached a maximum height (0.60) with around 500 classified grid cells, whereas $R^2$ reached 0.70 near the 1000-grid cell level for the WH field (Figure 3.11). Roughly 60% (in LD and RB) and 70% (WH) of the field variance in both cases was accounted for by making the clusters.

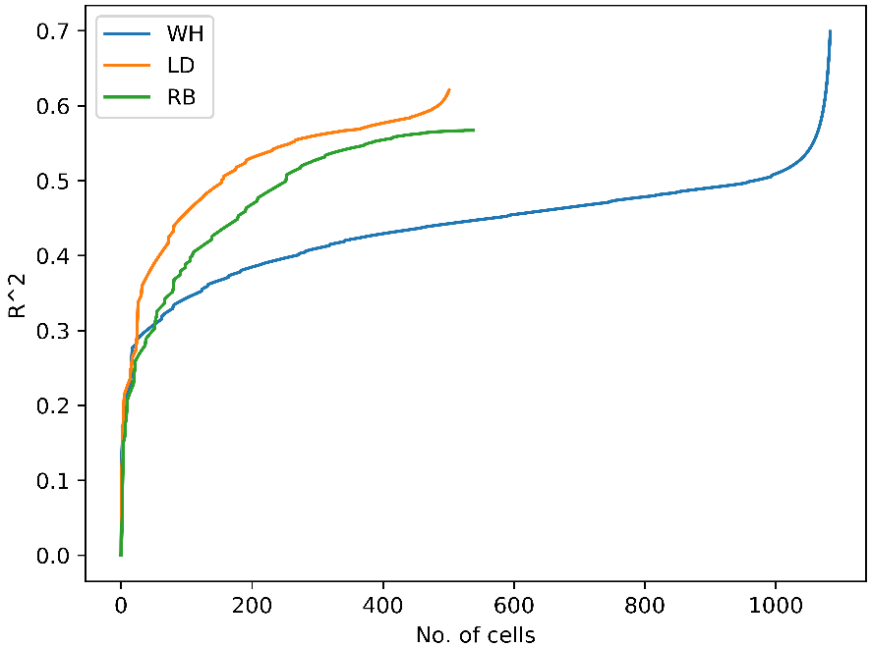



**Figure 3.11** Comparison of Coefficient of determination ($R^2$) value for Neighborhood Search Analyst (NSA) clustering for WH, LD, and RB fields.

### *3.3.4 Comparison of k-Means and NSA Clustering*

At this stage, three user-defined clusters (k = 5, k = 15, and k = 25) were generated to allow for a comparison of the two clustering algorithms, *i.e.*, k-means and NSA. User-defined centers for all clusters were needed for k-means; however, these were not a requirement for the NSA algorithm. The $R^2$ values of the NSA algorithm were compared among the three different fields (Figure 3.11). The overall OF showed that all of the clusters reached maximum $R^2$ values close to 0.6 and up to 0.7. In the three defined k-means clusters (k = 5, k = 15, and k = 25), the $R^2$ of the RB field was higher: 0.78, 0.80, and 0.84 respectively (Figure 3.12). Also, $R^2$ (k = 5) was relatively high in k-means clustering process because of the fragmentation of clusters throughout the field, while NSA clusters were always contiguous (*i.e.*, not broken into parts). The $R^2$ of the k-means cluster compared to that of the NSA was higher in most of the fields and was approximately 0.80. The $R^2$ values were comparable when the isolated/boundary pixels in each k-means cluster were disjointed from the main cluster and created spatially contiguous zones. The k-means cluster map consisted of groups or pixels with isolated boundaries in various parts of the WH field (Figure 3.7b), whereas the NSA algorithm counted these as different groups and reduced the zone

fragmentation (Figure 3.8a). In the case of the user-defined cluster (k = 25), the k-means cluster maps of WH, LD, and RB fields generated 36, 34, and 38 scattered zones respectively (Appendix B), whereas the NSA maps created approximately 25 spatially contiguous clusters for each of the three fields (Figure 3.12).

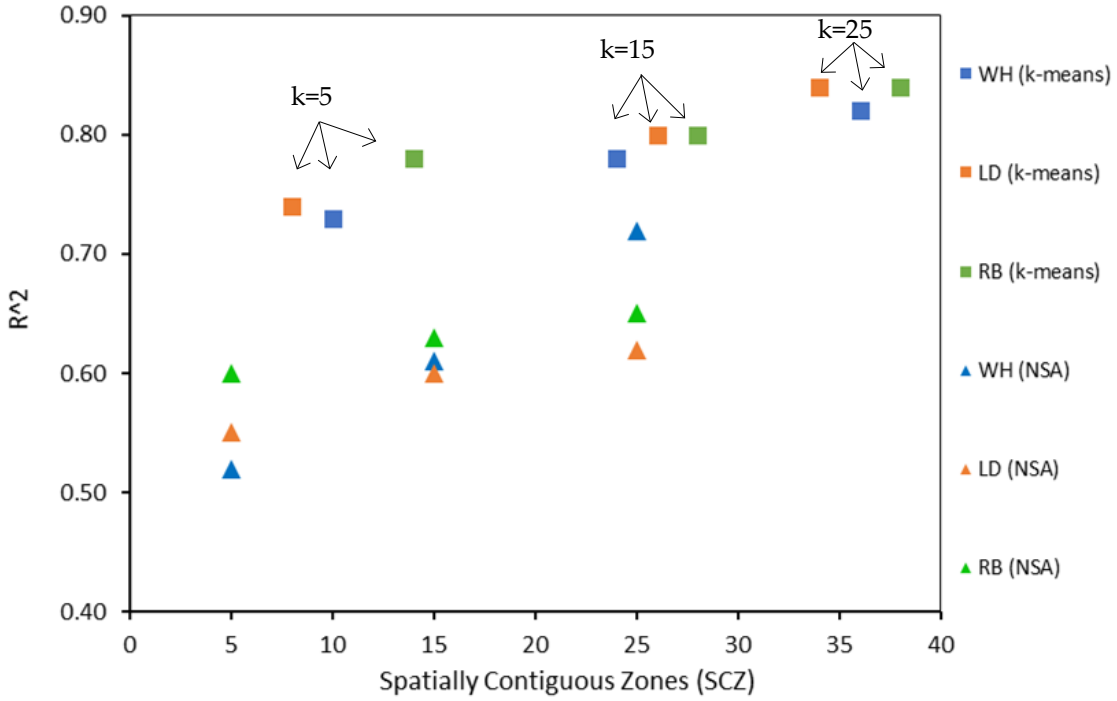



**Figure 3.12** Comparison of $R^2$ value between k-means and Neighborhood Search Analyst (NSA) clustering. The abscissa (SCZ) shows the number of spatially contiguous zones created when k = 5, k = 15, and k = 25.

### 3.4 Conclusions

The high-density and multivariate data clustering approach provided an optimal number of zones for three agricultural fields in Ontario, Canada. The preprocessing and variable selection steps common to all clustering techniques are imperative for providing a well-defined zonal boundary for developing management zones. Compared to other data clustering algorithms, NSA has a unique capability for zone separation, which allows one to produce an optimum number of zones and spatially contiguous clusters during multivariate classification. Moreover, an improved version of this software was tested and proved to be capable of handling a significant number of variables and data layers for delineating the optimum number of zones in a more robust way.

The software was found to be reliable when integrating high-density field topography, RS, and PSS data files. It had a fast processing time and could run on any platform with open source python modules. The robust zone delineation process and georeferenced thematic maps are useful for variable rate crop management technologies and for other management purposes. However, in order to optimize the field management strategies and verify management input across each zone, significant differences in the crop response would be an important parameter. Multi-sensor data fusion, advanced data filtering procedures, and the web application of the NSA could be implemented to facilitate the appropriate site-specific agronomic and environmental decisions in many regions.

The zonal maps will be useful for further agronomic model calibration using targeted soil sampling. Field data, for example, crop yield and lab-measured soil properties, could be used to validate the georeferenced clusters and management zones created.

**Funding**

Partial funding for this research was provided by Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) New Directions Research Program (ND2014-2487) and through the Graduate Merit Scholarship, Nature and Technology-FRQNT (B2X), Government of Quebec, Canada.

**Acknowledgments**

The authors are giving special thanks to the Woodrill Farms, Ontario, Canada, for the data support and cooperation. We would like to thank Nandkishor Dhawale, graduated from McGill University, for implementing the earlier version of the NSA algorithm in MATLAB. We are grateful to the Planet Labs for its free provision of Sentinel-2 data.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Zhang, N.; Wang, M.; Wang, N. Precision Agriculture—A Worldwide Overview. *Comput. Electron. Agric.* **2002**, *36*, 113–132.
2. Shatar, T.M.; McBratney, A. Subdividing a Field into Contiguous Management Zones Using a K-Zones Algorithm. In *3rd European Conference on Precision Agriculture*; Grenier, G., Blackmore, S., Ed.; Agro-Montpellier ENSAM: Montpellier, France, 2001; pp 115–120.
3. Fridgen, J.J.; Kitchen, N.R.; Sudduth, K.A; Drummond, S.T.; Wiebold, W.J.; Fraisse, C.W. Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation. *Agron. J.* **2004**, *96*, 100–108.
4. Khosla, R.; Westfall, D.G.; Reich, R.M.; Mahal, J.S.; Gangloff, W.J. Spatial Variation and Site-Specific Management Zones. In *Geostatistical Applications for Precision Agriculture*; Oliver, M.A., Ed.; Springer Science: Berlin, Germany, 2010; pp. 195–219.
5. De Benedetto, D.; Castrignano, A.; Diacono, M.; Rinaldi, M.; Ruggieri, S.; Tamborrino, R. Field Partition by Proximal and Remote Sensing Data Fusion. *Biosyst. Eng.* **2013**, *114*, 372–383.
6. Dhawale, N.M.; Adamchuk, V.I.; Prasher, S.O.; Dutilleul, P.R.L.; Ferguson, R.B. Spatially Constrained Geospatial Data Clustering for Multilayer Sensor-Based Measurements. In *Geospatial Theory, Processing, Modeling and Applications*; ISPRS Technical Commission II Symposium: Toronto, ON, Canada, 2014; Volume 40, pp. 187–190.
7. Castrignanò, A.; Buttafuoco, G.; Quarto, R.; Vitti, C.; Langella, G.; Terribile, F.; Venezia, A. A Combined Approach of Sensor Data Fusion and Multivariate Geostatistics for Delineation of Homogeneous Zones in an Agricultural Field. *Sensors (MDPI)* **2017**, *17*, 1–20.
8. Albornoz, E.M.; Kemerer, A.C.; Galarza, R.; Mastaglia, N.; Melchiori, R.; Martínez, C.E. Development and Evaluation of an Automatic Software for Management Zone Delineation. *Precis. Agric.* **2018**, *19*, 463–476.

9. Deng, X.; Wang, Y.; Peng, H. Clustering of High-Resolution Remote Sensing Imagery. In *Third International Asia-Pacific Environmental Remote Sensing Remote Sensing of the Atmosphere, Ocean, Environment, and Space*; Ungar, S., Mao, S., Yasuoka, Y., Eds.; SPIE: Hangzhou, China, 2003.

10. Adamchuk, V.I.; Hummel, J.W.; Morgan, M.T.; Upadhyaya, S.K. On-the-Go Soil Sensors for Precision Agriculture. *Comput. Electron. Agric.* **2004**, *44*, 71–91.

11. Berkhin, P. A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data*; Springer: Berlin/Heidelberg, Germany, 2006; pp 25–71.

12. Cohen, S.; Cohen, Y.; Alchanatis, V.; Levi, O. Combining Spectral and Spatial Information from Aerial Hyperspectral Images for Delineating Homogenous Management Zones. *Biosyst. Eng.* **2013**, *114*, 435–443.

13. De Benedetto, D.; Castrignanò, A.; Rinaldi, M.; Ruggieri, S.; Santoro, F.; Figorito, B.; Gualano, S.; Diacono, M.; Tamborrino, R. An Approach for Delineating Homogeneous Zones by Using Multi-Sensor Data. *Geoderma* **2013**, *199*, 117–127.

14. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* **2003**, *117*, 3–52.

15. Vrindts, E.; Mouazen, A.M.; Reyniers, M.; Maertens, K.; Maleki, M.R.; Ramon, H.; De Baerdemaeker, J. Management Zones Based on Correlation between Soil Compaction, Yield and Crop Data. *Biosyst. Eng.* **2005**, *92*, 419–428.

16. Yan, L.; Zhou, S.; Feng, L. Delineation of Site-Specific Management Zones Based on Temporal and Spatial Variability of Soil Electrical Conductivity. *Pedosphere* **2007**, *17*, 156–164.

17. Cressie, N.; Kang, E.L. High-Resolution Digital Soil Mapping: Kriging for Very Large Datasets. In *Proximal Soil Sensing*; Viscarra Rossel, R.A., McBratney, A.B., Minasny, B., Eds.; Springer: Dordrecht, The Netherlands, 2010; pp. 49–63.

18. Jiang, Q.; Fu, Q.; Wang, Z. Study on Delineation of Irrigation Management Zones Based on Management Zone Analyst Software. In *International Conference on Computer and Computing Technologies in Agriculture*; Springer: Berlin/Heidelberg, Germany, 2010; pp 419–427.

19. Adamchuk, V.I.; Viscarra Rossel, R.A. Precision Agriculture: Proximal Soil Sensing. In *Encyclopedia of Agrophysics*; Gliński, J., Horabik, J., Lipiec, J., Eds.; Springer: Dordrecht, The Netherlands, 2011; pp 650–656.

20. Dhawale, N., Adamchuk, V., Huang, H., Ji, W., Lauzon, S., Biswas, A., D.P. Integrated Analysis of Multilayer Proximal Soil Sensing Data. In Proceedings of the International Conference on Precision Agriculture, St. Louis, MO, USA, 31 July–4 August 2016.

21. Samet, H. An Overview of Hierarchical Spatial Data Structures. In Proceedings of the Fifth Israeli Symposium on Artificial Intelligence, Vision, and Pattern Recognition, Tel-Aviv, Ganei-Hata`arucha, Israel, 27-28 December 1988; pp. 331–351.

22. Arabie, P.; Hubert, L.J. An Overview of Combinatorial Data Analysis. In *Clustering and Classification*; Arabie, P., Soete, G.D., Hubert, L.J., Eds.; World Scientific Pub. Co.: Singapore, 1996; pp. 5–63.

23. Fisher, D. Iterative Optimization and Simplification of Hierarchical Clustering. *J. Artif. Intell. Res.* **1996**, *4*, 147–178.

24. Burrough, P.A.; Van Gaans, P.F.M.; Hootsmans, R. Continuous Classification in Soil Survey: Spatial Correlation, Confusion and Boundaries. *Geoderma* **1997**, *77*, 115–135.

25. Ruß, G.; Brenning, A. Data Mining in Precision Agriculture: Management of Spatial Information. In *Computational Intelligence for Knowledge-Based Systems Design*; Hüllermeier, E., Kruse, R., Hoffmann, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 350–359.

26. Johnson, S.C. Hierarchical Clustering Schemes. *Psychometrika* **1967**, *32*, 241–254.

27. Sadahiro, Y. Cluster Perception in the Distribution of Point Objects. *Cartogr. Int. J. Geogr. Inf. Geovisualization* **1997**, *34*, 49–62.

28. Fraisse, C.W.; Sudduth, K.A.; Kitchen, N.R. Delineation of Site-Specific Management Zones by Unsupervised Classification. *Trans. ASAE* **2001**, *44*, 155–166.

29. Motwani, M. A Study on Initial Centroids Selection for Partitional Clustering Algorithms. *Adv. Intell. Syst. Comput.* **2019**, *731*, 211–220.

30. De Gruijter, J.J.; Walvoort, D.J.J.; Van Gaans, P.F.M. Continuous Soil Maps—A Fuzzy Set Approach to Bridge the Gap between Aggregation Levels of Process and Distribution Models. *Geoderma* **1997**, *77*, 169–195.

31. Gui-Fen, C.; Li-Ying, C.; Guo-Wei, W.; Bao-Cheng, W.; Da-You, L.; Sheng-Sheng, W. Application of a Spatial Fuzzy Clustering Algorithm in Precision Fertilisation. *N. Z. J. Agric. Res.* **2007**, *50*, 1249–1254.

32. Panda, S.; Sahu, S.; Jena, P.; Chattopadhyay, S. Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study. *Adv. Intell. Soft Comput.* **2012**, *166*, 451–460.

33. Orhan, U.; Hekim, M.; Ozer, M. EEG Signals Classification Using the *K*-Means Clustering and a Multilayer Perceptron Neural Network Model. *Expert Syst. Appl.* **2011**, *38*, 13475–13481.

34. Saifuzzaman, M.; Adamchuk, V.; Huang, H.-H.; Ji, W.; Rabe, N.; Biswas, A. Data Clustering Tools for Understanding Spatial Heterogeneity in Crop Production by Integrating Proximal Soil Sensing and Remote Sensing Data. In Proceedings of the 14th International Conference on Precision Agriculture, Montreal, QC, Canada, 24–27 June 2018; p. 14. International Society of Precision Agriculture: Illinois, USA, Available online: http://www.ispag.org (Accessed on 20 June 2018).

35. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. Density-Based Clustering Algorithms for Discovering Clusters. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, 2–4 August 1996; pp. 226–231.

36. Liu, Y.; Xiong, N.; Zhao, Y.; Vasilakos, A.V.; Gao, J.; Jia, Y. Multi-Layer Clustering Routing Algorithm for Wireless Vehicular Sensor Networks. *IET Commun.* **2010**, *4*, 810.

37. Viscarra Rossel, R.A.; Adamchuk, V.I.; Sudduth, K.A.; McKenzie, N.J.; Lobsey, C. Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. *Adv. Agron.* **2011**, *113*, 237–283.

38. Córdoba, M.A.; Bruno, C.I.; Costa, J.L.; Peralta, N.R.; Balzarini, M.G. Protocol for Multivariate Homogeneous Zone Delineation in Precision Agriculture. *Biosyst. Eng.* **2016**, *143*, 95–107.

39. González-Fernández, A.B.; Rodríguez-Pérez, J.R.; Ablanedo, E.S.; Ordoñez, C. Vineyard Zone Delineation by Cluster Classification Based on Annual Grape and Vine Characteristics. *Precis. Agric.* **2017**, *18*, 525–573.

40. Lazarevic, A.; Xu, X.; Fiez, T.; Obradovic, Z. Clustering-Regression-Ordering Steps for Knowledge Discovery in Spatial Databases. In Proceedings of the International Joint Conference on Neural Networks Washington, DC, USA, 10–16 July 1999; 2530–2534.

41. Walters, R.W.; Jenq, R.R.; Hall, S.B. Evaluating Farmer Defined Management Zone Maps for Variable Rate Fertilizer Application. *Precis. Agric.* **2000**, *2*, 201–215.

42. Khosla, R., K.; Fleming, K.; Delgado, J.A.; Shaver, T.M.; Westfall, D.G. Use of Site-Specific Management Zones to Improve Nitrogen Management for Precision Agriculture. *J. Soil Water Conserv.* **2002**, *57*, 513–518.

43. Mondal, P.; Jain, M.; Defries, R.S.; Galford, G.L.; Small, C. Sensitivity of Crop Cover to Climate Variability: Insights from Two Indian Agro-Ecoregions. *J. Environ. Manag.* **2014**, *148*, 21–30.

44. Huang, Y.; Lan, Y.; Thomson, S.J.; Fang, A.; Hoffmann, W.C.; Lacey, R.E. Development of Soft Computing and Applications in Agricultural and Biological Engineering. *Comput. Electron. Agric.* **2010**, *71*, 107–127.

45. Beven, K.J.; Kirkby, M.J. A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrol. Sci. Bull.* **1979**, *24*, 43–69.

46. Roberts, D.F.; Adamchuk, V.I.; Shanahan, J.F.; Ferguson, R.B.; Schepers, J.S. Estimation of Surface Soil Organic Matter Using a Ground-Based Active Sensor and Aerial Imagery. *Precis. Agric.* **2011**, *12*, 82–102.

47. Viña, A.; Gitelson, A.A.; Nguy-robertson, A.L.; Peng, Y. Remote Sensing of Environment Comparison of Different Vegetation Indices for the Remote Assessment of Green Leaf Area Index of Crops. *Remote Sens. Environ.* **2011**, *115*, 3468–3478.

48. U.S. Department of Agriculture (USDA). *Management Zone Analyst Version 1.0 Software*; U.S. Department of Agriculture: Washington, DC, USA, 2000.

49. GNip, P.G.; Harvát, K.C. Management of Zones in Precision Farming. *Agric. Econ.* **2003**, *49*, 416–418.

50. Hartigan, J.A.; Wong, M.A. A K-Means Clustering Algorithm. *Appl. Stat.* **2012**, *28*, 100–108.

51. Nazeer, K.A.A.; Sebastian, M.P. Improving the Accuracy and Efficiency of the k-Means Clustering Algorithm. *Proc. World Cong. Eng.* **2009**, *I*, 1–5.

52. Vendrusculo, L.G.; Kaleita, A.F. Modeling Zone Management in Precision Agriculture through Fuzzy C-Means Technique at Spatial Database. In Proceedings of the Agricultural and Biosystems Engineering Conference, Louisville, KY, USA, 8–10 August 2011; Volume 4, pp. 2701–2715.

53. Bragato, G. Fuzzy Continuous Classification and Spatial Interpolation in Conventional Soil Survey for Soil Mapping of the Lower Piave Plain. *Geoderma* **2004**, *118*, 1–16.

54. Yan, L.; Zhou, S.; Feng, L.; Hong-Yi, L. Delineation of Site-Specific Management Zones Using Fuzzy Clustering Analysis in a Coastal Saline Land. *Comput. Electron. Agric.* **2007**, *56*, 174–186.

**Connecting Text to Chapter 4**

Chapter 4 is related to the second objective as listed in Chapter 1 as well as providing a rationale for using proximal soil sensing to estimate soil nutrients as discussed in Chapter 2. In the previous chapter, the optimization of various sensor data characterization was assessed for site-specific field management. It was shown that successful optimization of zonal variations could be determined by integrating PSS and RS density measurements. The findings demonstrated the improvement of the NSA hierarchical method in producing the optimum number of homogeneous zones compared to traditional clustering methods for field characterization. In Chapter 4, densely measured PSS data will be evaluated for field characterization, and it will be shown to optimize the efficiency of soil prediction. The spatial data clustering algorithm is evaluated as a calibration sampling design tool to investigate the effectiveness of the prediction model for estimating multiple soil properties. As most PSS systems do not directly measure soil nutrients, they require a further calibration procedure to relate sensing measurements to estimate multiple soil properties. Therefore, an evaluation of lab-based soil sample analysis is performed for assessing data quality and prediction efficiency. Then, the prediction results are validated through the reported error of estimation with North American lab-based results for improving the quality of the prediction.

Initial outcomes were reported and published at a professional society meeting, conference proceedings and a journal:

1. Ji, W., Adamchuk, V., Lauzon, S., Su, Y., Saifuzzaman, M., & Huang, H. (2017). Pre-processing of on-the-go mapping data. In *The Book of Abstracts for Pedometrics 2017 Conference, 26 June - 1 July 2017* (p. 113). Wageningen, the Netherlands.
2. Saifuzzaman, M., Adamchuk, V., Biswas, A. & Dutilleul, P. R. L. (2019). Soil Prediction using High-Density Data for Understanding Field Variability and Crop Management. In *Abstracts from Annual Meeting of the Association of American Geographers, April 3 - 7 2019*. Washington DC, USA: (CD publication).
3. Saifuzzaman, M., Adamchuk, V., Biswas, A., and Rabe, N. (2020). High-density Proximal Soil Sensing Data and Topographic Derivatives to Characterize Field Variability. *Biosystems Engineering - Elsevier* (In preparation).

# Chapter 4: High-density Proximal Soil Sensing Data and Topographic Derivatives to Characterize Field Variability

Md Saifuzzaman, Viacheslav Adamchuk, Asim Biswas, and Nicole Rabe

**Abstract**

Proximal soil sensing platforms can provide high-density yet affordable sensor data to describe agricultural field variability. The availability of such data, along with recent advances in analysis methods, allows for the optimization of model errors and a determination of their spatial variability. Most current sensors measure field parameters indirectly, rather than directly linking them to agronomic properties relevant to crop growth. Uncertainty analysis for predicting soil properties is an emerging challenge in precision agricultural practice. High-density soil sensor data and their capacity to contribute to the prediction of soil properties was investigated. An assessment of model accuracy was made by comparing model outputs to validation data points. High-accuracy topography and apparent soil electrical conductivity ($EC_a$) mapped with either DUALEM-21S or RTK GNSS sensors were used to characterize field-scale soil variability at 13 field sites in Ontario. Lab analyses of six soil properties [pH; buffer pH (BpH); Soil Organic Matter (SOM); Phosphorus (P); Potassium (K); and Cation Exchange Capacity (CEC)] were undertaken to characterize soil variability across the fields. DUALEM-21S sensor variables were co-linear to one another. The topographic variables of slope and topographic wetness index, along with the remainder of the sensor variables, were key inputs to the prediction model. High Pearson's correlation coefficients ($r \geq 0.60$) indicated strong correlations between sensor variables and field-measured soil properties, topographic parameters and shallow $EC_a$ (PRP1: 0 – 0.5 m) sensor variables, allowing effective predictions of several chemical properties (*i.e.*, SOM, P, and CEC) at different locations. Among the 13 agricultural fields, two fields presented the best-structured data, resulting in the lowest prediction errors. Drawing on topographic variables provided promising predictions of field SOM and CEC. This highlights the powerful potential of proximal soil sensing technologies to define the site-specific crop production environment in terms of terrain and physical characterization of the soil. The integration of conceptually different sensors allows for better prediction of certain soil properties than a single measurement approach.

**Keywords:** Proximal soil sensing, Topographic derivatives, Soil properties, Error estimation, validation.

## 4.1 Introduction

In this age of precision farming, crop scientists draw upon densely measured surface and subsurface information to assess soil distribution patterns and crop nutritional requirements (Adamchuk *et al.*, 2004; Lück *et al.*, 2009; Alchanatis and Cohen, 2013; Pierpaoli *et al.*, 2013). To achieve site-specific management across a landscape, they also consider the local soil-crop relationship and its variability. However, when determining the most economical local fertilization needs, the high cost of laboratory soil analysis limits the conventional means of characterizing variability (Huang *et al.*, 2014). To fulfill current demand, recent agricultural technologies have proven effective at collecting high-density soil information (Friedman, 2005; Viscarra Rossel *et al.*, 2011; Walker *et al.*, 2017) by drawing upon remote and proximal soil sensing (PSS) technologies (Alchanatis and Cohen, 2013; Viscarra Rossel and Adamchuk, 2013; Aldabaa *et al.*, 2015). With high-density data, new PSS technologies have facilitated the delineation of the spatial distribution of soil edaphic properties across agricultural fields in North America (Adamchuk and Tremblay, 2017). Long processing times for collected soil samples and concerns regarding local-scale precision have led to large grain producers relying on spatial and temporal surface and sub-surface soil sensor information (Zhang *et al.*, 2002; Kerry *et al.*, 2017).

Mounted on a range of sensing platforms, various proximal soil sensing technologies are being developed to provide high-density, yet affordable, data, providing a detailed representation of field heterogeneity. In the past few decades, passive and active PSS systems have contributed to our understanding of the soil-topography relationship and assessed spatial variability for precision farming (Brown, 2006; Rodrigues *et al.*, 2015; Neely *et al.*, 2016; Hutengs *et al.*, 2019). Data from geoelectrical and electromagnetic sensors are widely used for identifying soil dielectric properties and geospatial variability (Adamchuk and Viscarra Rossel, 2010; Singh *et al.*, 2016; Watson *et al.*, 2017). In many regions in Canada, electromagnetic-energy-enabled on-the-go PSS sensors (*e.g.*, DUALEM-21S, EM-38, etc.) have served to inform soil management practices under precision farming. However, most sensors document changes in parameters that indirectly, rather than directly, affect agronomic indicators of the crop growing environment (Adamchuk *et al.*, 2005; Vitharana *et al.*, 2008). Corwin and Lesch (2005) and Friedman (2005) showed that the precise locations and the depths of apparent electrical conductivity ($EC_a$) measurements are highly correlated to top- and sub-soil physical properties (*e.g.*, depth of clay layer, soil salinity, and water

content). However, the $EC_a$ data collected from variable depths is required to further process site-specific depth exploration before linking the measurements to soil constituents (Saey *et al.*, 2009; Sun *et al.*, 2011; Stockmann *et al.*, 2017; Zare *et al.*, 2018; Nocco *et al.*, 2019).

High-density data is a necessary element of digital soil mapping and for making agro-economic decisions (McBratney *et al.*, 2003). Widely implemented on a single platform, Real-time kinematic (RTK) global navigation satellite systems (GNSS) are combined with other sensors to construct dense georeferenced maps of surface topography. The digital terrain model from the georeferenced points serve as a predictor of topographic variables in predicting soil attributes (Bishop and Minasny, 2006). Many derivatives [*e.g.*, Topographic wetness index (TWI); slope, aspect, etc.] from the terrain model are used in assessing topographic diversity, water movement, and water holding capacity as they relate to crop growth (Odeha *et al.*, 1994; Demattê *et al.*, 2006; Miller *et al.*, 2015). Along with the topographic variables, dense $EC_a$ measurements can also predict the presence and states of primary and secondary soil nutrients (Taylor *et al.*, 2003; Adamchuk and Viscarra Rossel, 2011; Dao, 2017). The georeferenced locations, lab-measured soil properties, and other corresponding sensor measurements can then be used to make management decisions for agricultural fields.

Geospatial and geostatistical analyses of different sensor variables and predictive approaches are key to developing management tools employed in precision farming (Adamchuk and Viscarra Rossel, 2010; Hengl *et al.*, 2017). Using dense georeferenced measurements to achieve a precise agricultural management solution involves data processing tools, approaches and models. Sun *et al.* (2011) and Viscarra Rossel *et al.* (2011) found that using a variable data structure for different PSS measurements improved the accuracy of prediction for soil properties, thereby, providing additional information for thematic mapping. The relationship among different available sensor variables are important in the data mining and decision-making processes. Multivariate statistical methods (*e.g.*, correlation and regression, principal component analysis, and semi-variograms) are commonly used for data preprocessing and structure analysis (McBratney *et al.*, 2000; Córdoba *et al.*, 2013). Accordingly, multivariate regression analysis has become a popular approach for soil characterization and the prediction of macro- or micro-nutrients.

Uncertainty analysis of the prediction model is an emerging challenge in precision soil mapping (Bishop and Minasny, 2006; Viscarra Rossel *et al.*, 2016; Duda *et al.*, 2017). To quantify model accuracy, various statistical tests are performed, including comparisons of mean squared error (MSE) values relative to the validation points. In previous studies, model sensitivity and errors were reported by different methods and minimized through different procedures (Oliver, 2010; Sudduth *et al.*, 2013). In the present study, the ratio of the standard error (SE) of prediction to the standard deviation (STD) of the sample serves to assess the model's performance. Moreover, to explain the proportion of variations in the regression line of the estimates, the adjusted coefficient of determination (adjusted $R^2$) is also reported. When the propagated SE of the estimate is optimum compared to the sample SD, the models are recommended for thematic mapping and soil management in precision agriculture (Adamchuk and Viscarra Rossel, 2011; Manasny and McBratney, 2013; Panayi *et al.*, 2017). The results must be validated with lab-based measurements. Recent developments in error modeling requirements may integrate high-density data points from various sensors to identify comprehensive soil nutrients and their distribution patterns on various geospatial scales (Zhou *et al.*, 2016; Castrignanò *et al.*, 2017; Minasny and McBratney, 2016). In terms of data optimization, the present study provides error-handling methods for PSS measurements.

In an overall effort to efficiently interpret the results from high-density data, the main goal of this research was to assess soil sensor data and its predictive capacity by evaluating various soil mapping techniques for various soil properties. This study was designed to assess proximal soil sensing-based predictability of physical and chemical soil attributes for a series of Ontario fields. The prediction results were validated by comparing them to the average values of North American lab-based soil measurements. This research takes a further step toward achieving a better understanding of both the advantages and limitations of contemporary proximal soil sensing solutions.

## 4.2 Materials and methods

### *4.2.1 Experimental fields*

Thirteen production fields across southeastern Ontario, Canada, differing in size and agro-climatic conditions, were selected for this study. They had their topography and soil mapped by

one of two popular commercial proximal soil sensing services (RTK GNSS or DUALEM-21S); they were then sampled manually at locations based on a neighbourhood search analyst clustering of the proximal-sensing outputs and tested in the laboratory (Table 4.1 & Figure 4.1). According to US soil taxonomy classes, the soil orders of the region are Alfisols and Spodosols (Luvisolic and Brunisolic, respectively, in Canadian soil order). Soil textural classes varied from sandy to clay loam. Based on soil survey data and a soil map of Ontario, KM and RL fields were in a typical Grenville soil association, a strongly calcareous and sandy loam texture with low moisture retention (OMAFRA, 2016). Among the seven Canada Land Inventory (CLI) land classes, all study sites were highly capable (Class 1 to 3) of supporting agriculture and land use activities. Mostly located in the northern and southwestern parts of the Lake Ontario watershed and influenced by the surrounding Great Lakes, the fields were under a humid continental climate. Overall, the fields were well managed in terms of runoff and drainage conditions (according to soil texture) for cropping. Elevation varied from a few meters to a hundred meters between the fields. In addition to the differences in elevation and drainage conditions, the study sites had good crop production histories, with corn (*Zea mays* L.) and soybean [*Glycine max* (L.) Merr.] as the main crops (Table 4.1).

**Table 4.1** Characteristics of thirteen agricultural fields in Ontario, Canada, including their area, number of soil samples, soil type, natural drainage conditions and primary crops.

| Field ID | Area (ha) | Number of samples | Soil Texture Class* | Drainage condition | Target crops** |
|---|---|---|---|---|---|
| F25 | 26 | 26 | Clay loam | Good | Wheat |
| WH | 40 | 99 | Loam | Very good | Soybean/Corn |
| KM | 30 | 119 | Silty loam | Poor | Soybean/Corn |
| LP | 34 | 72 | Clay loam | Good | Soybean |
| LD | 21 | 62 | Sandy loam | Very good | Corn/Wheat |
| TE | 39 | 97 | Sandy loam | Very good | Soybean |
| SM | 28 | 74 | Clay loam | Good | Soybean |
| NX | 48 | 74 | Clay | Poor | Corn/Soybean |
| R50 | 51 | 51 | Clay loam | Good | Wheat |
| RB | 75 | 72 | Fine Sandy loam | Very good | Soybean/Corn |
| RL | 47 | 49 | Sandy loam | Very good | Corn |
| ST | 39 | 76 | Clay | Poor | Corn/Soybean |
| VN | 20 | 51 | Silty loam | Poor | Wheat/Soybean |

** Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA) 2016 and 2017.*
*** OMAFRA 2017 and Grain Farmers Ontario (GFO)*


**Figure 4.1** Location of the thirteen agricultural fields under study in Ontario, Canada.

### *4.2.2 Soil sensing by proximal soil sensors*

A vehicle equipped with two types of proximal soil sensors (RTK GNSS and DUALEM-21S) was used for topographic and soil mapping between 2014 and 2017. The data from both sensors were logged using custom DUALEM_DAQ data acquisition software. Despite diverse data sources and a lack of standardization, generic rules were developed in terms of data format and preprocessing steps to assess the PSS data sensitivity to bare soil properties (Ji *el al.*, 2017). Timestamps, locations, speed of the sensor vehicle, the distance between data points, and other variable measurements were evaluated in the preprocessing steps. Various procedures were considered: (*i*) median filtering of neighboring measurements, and (*ii*) removing outliers (start- and end-pass delays, over speed limits, measurements outside the acceptable limit). Potential outliers and null values of the PSS measurements were identified in this step, and about 12% of

the PSS data was removed. Figure 4.2 shows the methodological development of this research. Once the sensor data were collected from the fields, soil samples were collected for laboratory analysis based on a field variability map derived from the sensor response. Various statistical analyses were performed on the sensor data and laboratory results. Finally, the variability analysis of different laboratory results and their predictive capability using field sensor data were assessed in comparison with the North American Proficiency Testing (NAPT) lab results.

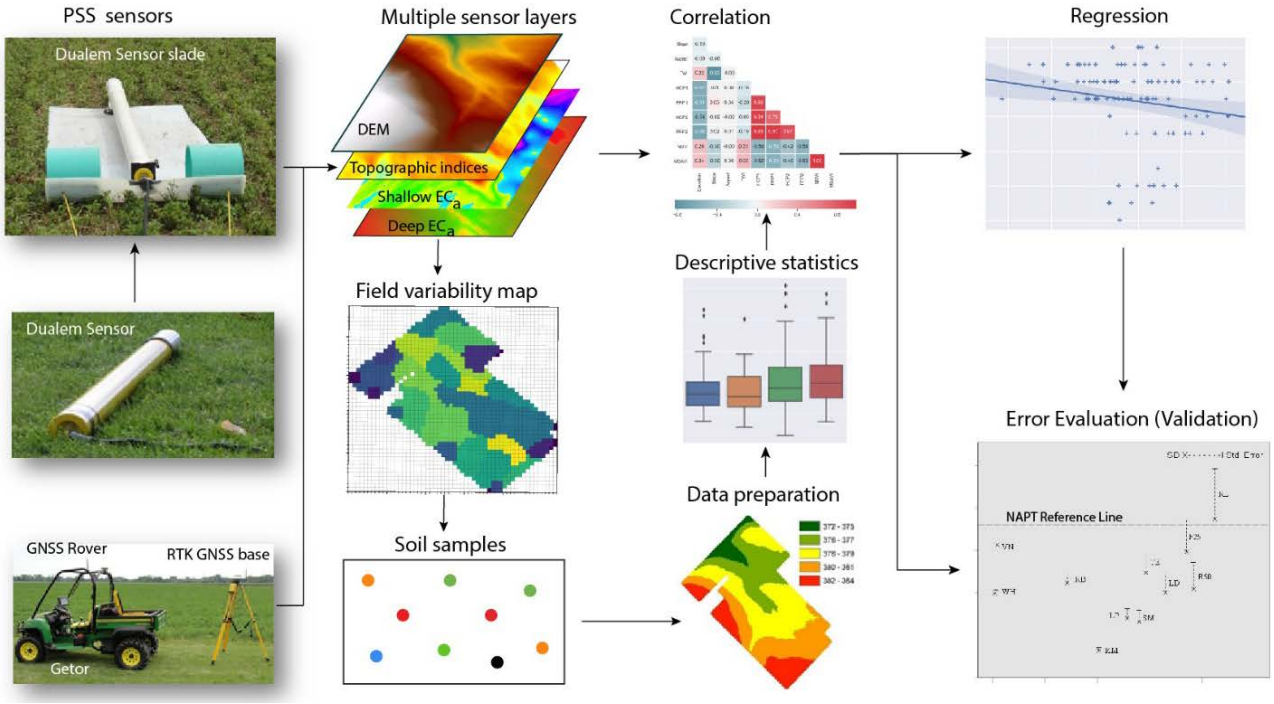



**Figure 4.2** Flow chart shows the research methods towards the error evaluation and validation.

*4.2.2.1 Soil sensing – $EC_a$ measurements*

Apparent soil electrical conductivity ($EC_a$) was obtained using an electromagnetic induction (DUALEM-21S instrument Inc., Milton, ON, Canada) method. The instrument (with two-pairs of electromagnetic receivers: horizontal co-planar geometry-HCP and perpendicular geometry-PRP) was used to collect soil apparent electrical conductivity ($EC_a$) at four different depths: HCP1 – 0-1.6 m ($EC_a^{0\text{-}1.6}$), PRP1 – 0-0.5 m ($EC_a^{0\text{-}0.5}$), HCP2 – 0-3.2 m ($EC_a^{0\text{-}3.2}$), and PRP2 – 0-1.0 m ($EC_a^{0\text{-}1.0}$) (Table 4.2 and 4.3). Descriptive statistics [Minimum (Min); median; Maximum (Max); Standard deviation (STD); and mean] were calculated for data assessment. Due to its high STD and Max values (Table 4.2 and 4.3), the ST field was not considered for further analysis.

Generally, high collinearity was found among $EC_a$ variables, although these were measured for different depths (Ji *et al.*, 2019).

**Table 4.2** DUALEM-21S sensor (HCP1 & PRP1) data collected from 13 agriculture fields located in Ontario, Canada.

| Field ID | # of measurements | HCP1($EC_a^{0.1.6}$) | | | | | PRP1 ($EC_a^{0.0.5}$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Min | Median | Max | STD | Mean | Min | Median | Max | STD | Mean |
| F25 | 2614 | 7.72 | 20.43 | 61.60 | 3.62 | 20.70 | 4.04 | 13.47 | 47.80 | 3.13 | 13.59 |
| WH | 20129 | 4.00 | 12.28 | 25.28 | 1.69 | 12.51 | 4.68 | 7.92 | 22.24 | 1.60 | 8.15 |
| KM | 11427 | 16.82 | 27.90 | 38.96 | 2.27 | 28.09 | 9.84 | 15.84 | 34.58 | 2.70 | 16.10 |
| LP | 7373 | 12.04 | 16.78 | 29.94 | 2.34 | 17.29 | 7.28 | 11.62 | 24.82 | 2.40 | 12.17 |
| LD | 6931 | 2.58 | 6.90 | 16.08 | 1.55 | 6.96 | 0.72 | 4.44 | 14.12 | 1.38 | 4.55 |
| TE | 11111 | 4.64 | 9.22 | 51.14 | 3.94 | 10.05 | 2.94 | 6.12 | 18.48 | 2.14 | 6.61 |
| SM | 7473 | 8.90 | 20.22 | 30.66 | 2.85 | 20.45 | 7.88 | 15.60 | 27.36 | 3.15 | 15.82 |
| NX | 4472 | 1.77 | 24.48 | 47.71 | 5.20 | 23.53 | 1.57 | 16.36 | 28.30 | 4.32 | 15.67 |
| R50 | 4659 | 11.70 | 19.63 | 37.21 | 3.05 | 19.87 | 5.10 | 11.74 | 25.00 | 2.78 | 11.91 |
| RB | 18524 | 1.70 | 9.00 | 17.98 | 2.81 | 9.13 | 0.00 | 3.53 | 16.80 | 2.86 | 4.40 |
| RL | 5898 | 3.20 | 7.66 | 43.02 | 2.96 | 8.35 | 0.02 | 2.06 | 30.52 | 2.43 | 2.90 |
| ST | 9337 | 1.00 | 26.62 | 110.28 | 13.06 | 26.27 | 1.28 | 16.68 | 107.84 | 9.56 | 17.62 |
| VN | 5073 | 8.56 | 34.60 | 65.82 | 9.26 | 32.97 | 5.02 | 23.64 | 43.36 | 6.30 | 22.84 |

**Table 4.3** DUALEM-21S sensor (HCP2 & PRP2) data was collected from 13 agriculture fields located in Ontario, Canada.

| Field ID | # of measurements | HCP2 ($EC_a^{0.3.2}$) | | | | | PRP2 ($EC_a^{0.1.0}$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Min | Median | Max | STD | Mean | Min | Median | Max | STD | Mean |
| F25 | 2614 | 13.42 | 19.37 | 54.60 | 3.14 | 19.75 | 7.92 | 17.96 | 59.00 | 3.66 | 18.20 |
| WH | 20129 | 7.42 | 10.46 | 24.42 | 1.79 | 10.83 | 5.42 | 9.10 | 23.92 | 1.75 | 9.37 |
| KM | 11427 | 24.58 | 29.66 | 39.22 | 1.51 | 29.67 | 17.30 | 23.98 | 38.96 | 2.79 | 24.21 |
| LP | 7373 | 6.92 | 13.00 | 25.84 | 1.87 | 13.30 | 8.88 | 14.22 | 29.52 | 2.52 | 14.80 |
| LD | 6931 | 0.50 | 4.44 | 14.44 | 1.85 | 4.61 | 1.08 | 4.68 | 14.60 | 1.50 | 4.75 |
| TE | 11111 | 2.02 | 5.56 | 87.54 | 6.78 | 6.86 | 2.33 | 6.56 | 19.54 | 2.27 | 7.09 |
| SM | 7473 | 6.12 | 14.80 | 23.04 | 2.20 | 15.05 | 8.28 | 17.68 | 28.86 | 3.05 | 17.87 |
| NX | 4472 | 1.75 | 22.80 | 57.80 | 5.73 | 21.91 | 1.65 | 20.96 | 41.75 | 5.11 | 20.25 |
| R50 | 4659 | 11.00 | 18.80 | 55.98 | 3.33 | 19.27 | 8.40 | 16.32 | 31.14 | 3.07 | 16.50 |
| RB | 18524 | 2.50 | 8.45 | 14.99 | 2.65 | 8.22 | 0.14 | 5.10 | 15.00 | 2.96 | 5.64 |
| RL | 5898 | 3.68 | 8.20 | 27.18 | 2.81 | 8.41 | 1.40 | 4.00 | 42.46 | 3.09 | 4.85 |
| ST | 9337 | 1.48 | 26.66 | 102.22 | 14.34 | 25.29 | 2.02 | 23.18 | 122.42 | 12.49 | 23.61 |
| VN | 5073 | 4.80 | 31.36 | 72.58 | 10.20 | 29.68 | 5.88 | 31.74 | 62.04 | 8.88 | 30.19 |

*4.2.2.2 Soil survey – RTK topographic mapping*

Topographic data for the agricultural fields were collected using a Trimble Real-Time Kinematic (Trimble Inc., Sunnyvale, California, USA) operating on a Global Navigation Satellite System (Table 4.4). Maximum individual field elevations ranged from 40 m to 380 m. Slope and aspect ratio [$\sin\frac{aspect}{2}$] ranges were derived from maximum/minimum elevations, while the topographic wetness index (TWI) was derived from a digital elevation model (DEM) of the study sites. Besides ArcGIS v10.5 (ESRI, Redlands, California, USA) software used in the geospatial analysis of topographic variables, the SAGA GIS system v. 6.3.0 (Departments of Physical Geography, Hamburg and Göttingen, Germany) software tool was used for calculating TWI. Among the twelve agricultural fields, F25, KM, NX, R50, ST, and VN had negligible elevation and gradient differences.

**Table 4.4** Summary statistics of elevation from RTK in 13 agricultural fields located in Ontario, Canada.

| Field ID | # of measurements | Elevation (m) | | | | |
|---|---|---|---|---|---|---|
| | | Min | Median | Max | STD | Mean |
| F25 | 12778 | 336.01 | 337.29 | 338.64 | 0.61 | 337.35 |
| WH | 28493 | 372.06 | 378.07 | 384.54 | 2.33 | 378.21 |
| KM | 11662 | 36.71 | 38.57 | 39.11 | 0.16 | 38.56 |
| LP | 7559 | 263.88 | 269.72 | 273.85 | 1.92 | 269.41 |
| LD | 7110 | 332.70 | 344.86 | 354.17 | 5.76 | 343.95 |
| TE | 17628 | 298.49 | 307.50 | 311.87 | 3.05 | 307.02 |
| SM | 7603 | 263.69 | 266.58 | 273.67 | 1.60 | 266.76 |
| NX | 4375 | 63.28 | 64.01 | 68.29 | 0.60 | 64.20 |
| R50 | 18326 | 330.24 | 331.58 | 333.24 | 0.65 | 331.48 |
| RB | 20813 | 358.41 | 367.67 | 372.16 | 3.63 | 366.64 |
| RL | 8230 | 185.56 | 194.70 | 222.69 | 3.31 | 194.47 |
| ST | 9429 | 57.06 | 59.71 | 66.74 | 2.09 | 60.18 |
| VN | 5181 | 31.80 | 38.40 | 46.70 | 2.21 | 38.50 |

*4.2.3 Soil sampling and laboratory analysis*

Based on the RTK and DUALEM-21S sensor measurements in the agricultural fields, soil samples were collected for laboratory analysis. The Neighborhood search analyst (NSA) data clustering tool developed in previous study (See Chapter 3) which implemented an optimization of calibration sample placement was applied to locate soil sampling points for developing site-

specific soil property maps (Saifuzzaman *et al.*, 2019). Based on the maximum field variability analyzed from the previously collected PSS measurements, the NSA algorithm determined cluster centers and the optimum number of sample sizes given the field's acreage. 1-acre grid-based sampling strategy was also applied in some fields. Based on the two different methods, sampling density varied across the study sites (Table 4.1). Sampling points were positioned in the fields using a Garmin handheld wide area augmentation system-corrected GPS and georeferenced. Soil samples were collected from the sites at the beginning of the cropping seasons. The lab-measured soil samples were processed, and specific parameters selected for the prediction model. In this study, the six major lab-measured soil properties targeted were pH, buffer pH (BpH), Soil Organic Matter (SOM), plant available Phosphorus (P) and Potassium (K), and Cation Exchange Capacity (CEC). BpH data was important for the soil analysis and only available for fields RL, KM, LD, NX, and VN. Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA) accredited soil test methods were used for analyzing all field samples: pH – 1:2 saturated paste; BpH – SMP Buffer solution; OM% - Walkley-Black (0-8%), Loss on Ignition (>8%); P – Olsen sodium bicarb; K – ammonium acetate extract; and CEC – calculated by converting soil test K/Mg/Ca to milliequivalents. Those major soil properties determine fertilizer and lime requirements for site-specific soil and crops.

All predictor variables were independently assessed based on the characteristic of the variable (Table 4.5), then prepared for spatial interpolation and soil prediction. Topographic variables were assessed based on the terrain attributes collected from RTK GNSS. Four $EC_a$ variables were assessed with general statistical methods.

**Table 4.5** The description of the sensor variables and the measured soil properties.

| Sources (Sensor/Lab) | Target Covariables | Characteristics of the Variable |
|---|---|---|
| RTK GNSS | Terrain attributes | Elevation (m)<br>Topographic wetness index -TWI<br>Slope (%)<br>Aspect ratio |
| DUALEM-21S | Soil $EC_a$ (mS $m^{-1}$) | HCP1 (0-1.6 m)<br>HCP2 (0-3.2 m)<br>PRP1 (0-0.5 m)<br>PRP (0-1.0 m) |
| Soil sample properties: (Lab analysis) | pH | 0 to 14 |
| | BpH | 0 to 14 |
| | SOM | Soil organic matter (% w/w) |
| | P | Soil Phosphorus (ppm) |
| | K | Soil Potassium (ppm) |
| | CEC | Cation exchange capacity (meq $hg^{-1}$) |

*4.2.4 Spatial interpolation and point data extraction*

Interpolation using spatial autocorrelation was performed to delineate the variability of point-based PSS data and spatial characteristics (García-Tomillo *et al.*, 2016). Ordinary Kriging interpolation maps were generated from topographic (Figure 4.3) and $EC_a$ measurements in ESRI ArcGIS software (v10.5). Various geospatial (*e.g.*, rectification, point data extraction, etc.) tools were used for data processing and further analysis. Multiple kriged maps delivered spatial covariates associated with sampling points in a data file (Table 4.5). Finally, the text data file containing multiple layers of sensor variables and soil measurements of each study site was used for statistical analysis and error mapping in an open source data analysis platform (Python Pandas).

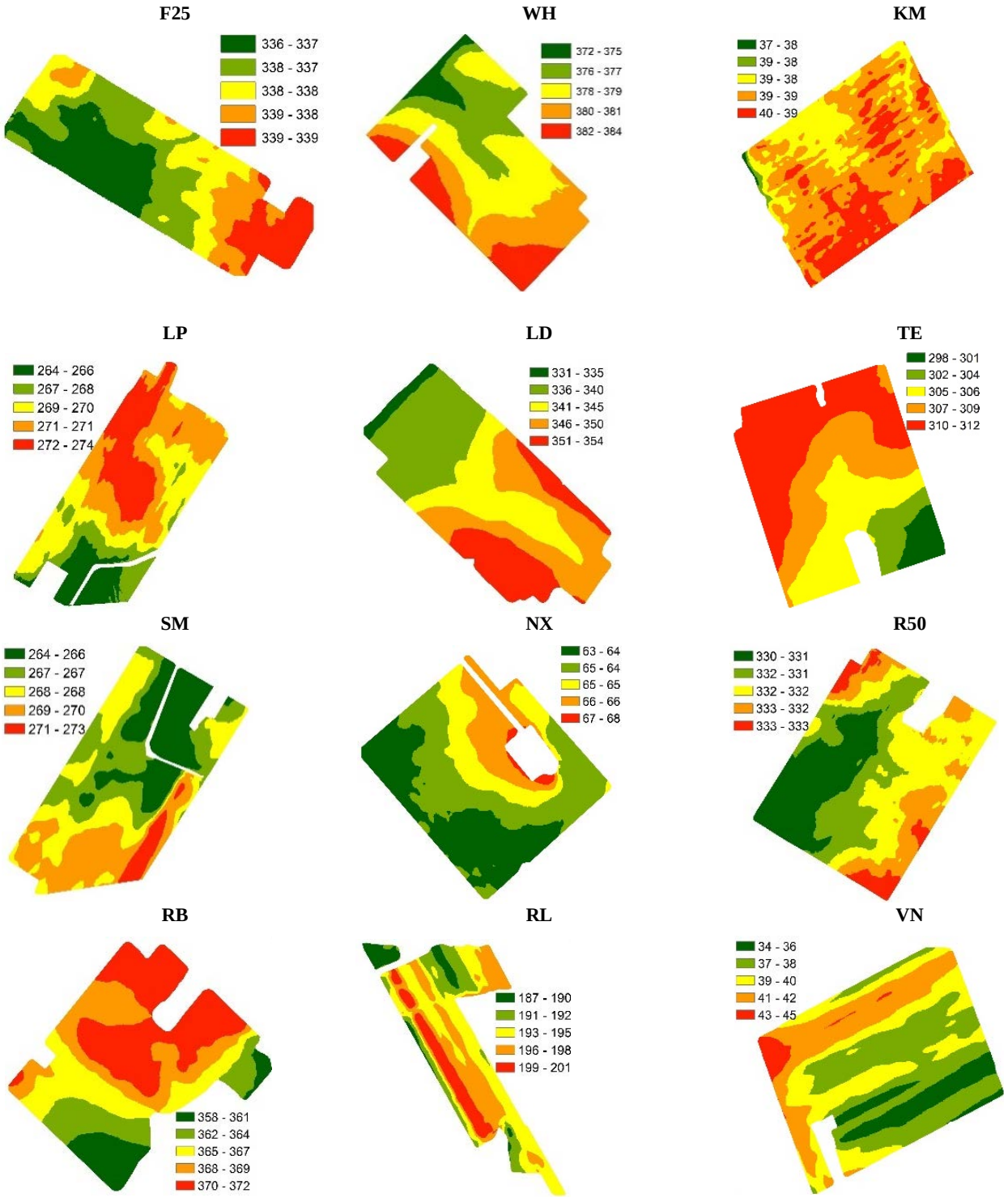



**Figure 4.3** The interpolated elevation (in meters) maps, showing field variability in the twelve study sites.

*4.2.5 Correlation and regression analysis*

Multicollinearity was used to assess the spatial data correlation among the predictor variables (sensor variables and sensor-derived variables). High collinearity was found mostly among $EC_a$ variables. Slope and TWI were not correlated with the remaining variables. Given the relationship between sensor data and soil measurements, the Pearson's correlation ($r$) was assessed as:

$$r = \frac{\sum_{i=1}^{i=n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{i=n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{i=n}(y_i - \bar{y})^2}} \quad (1)$$

where, $n$ is the total number of measurements, $x_i$ and $y_i$ are the $i^{th}$ individual values of variables $x$ and $y$, and $\overline{x}$ and $\overline{y}$ are the means of variables $x$ and $y$, where the $x$ values represent soil measurements, and $y$ values are derived from sensor data. To further evaluate the linear relationship and prediction error between sensor measurements and measured soil variability, the ordinary least square (OLS) was employed. Excel's Regression tool (Data Analysis in Microsoft Excel 2016) was used for the regression analysis.

*4.2.6 Error estimation in model prediction*

To determine the prediction error, statistical parameters for the samples' sensor variables and their associated soil parameter estimates were derived. The error was evaluated by the $s$ of the sample measurements and an estimation of the standard error of the mean ($s_{\bar{x}}$). The ratio of the $s_{\bar{x}}$ in predicting the $s$ of the samples was assessed and served in scaling the level of error for the soil prediction model. Moreover, the coefficient of determination (adjusted $R^2$) was used to explain the proportion of variation in the estimates and to evaluate the model's predictive performance. When the difference of $s_{\bar{x}}$ of the estimate and the sample standard deviation ($s$) is smaller, and the adjusted $R^2$ values are considerably greater among the study sites, the data can make accurate predictions, and therefore, can be recommended.

For validation, the reported errors were compared to the soil analysis results published online by the North American Proficiency Testing (NAPT) program. The median absolute deviation (MAD) of the NAPT results contributes to the continuous improvement and heightened precision of the analytical results for agricultural soils throughout North America (NAPT, 2019).

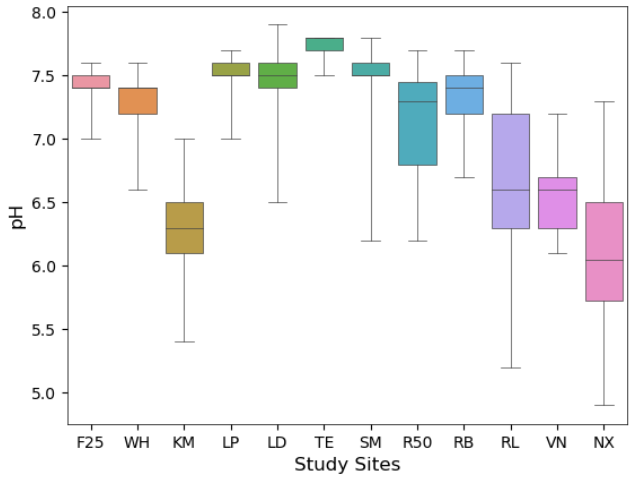
In the present study, a reference line (value) was calculated from the MAD of the NAPT results by averaging the values of the last ten years (2009 to 2019). When the error estimation ($s$ and $s_{\bar{x}}$) is below the average line of the NAPT values among the study sites, the data can make relatively precise predictions.

## 4.3 Results

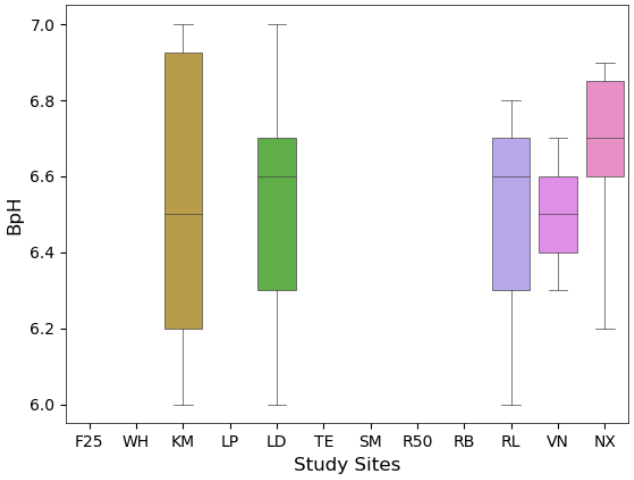
### *4.3.1 Descriptive statistics*

High-density $EC_a$ measurements (at four different depths: HCP1, HCP2, PRP1, and PRP2) were assessed through descriptive statistics (Table 4.2 and 4.3). Across the twelve sites, the range of $EC_a$, $0.02 \leq EC_a^{0\cdot0.5} \leq 47.80$, $1.70 \leq EC_a^{0\cdot1.6} \leq 65.82$, $0.14 \leq EC_a^{0\cdot1.0} \leq 62.04$, and $0.50 \leq EC_a^{0\cdot3.2} \leq 87.54$, were determined and showed large variability. In terms of the $EC_a$ sensor measurements, spatially close (Figure 4.1) and topographically (Table 4.4) similar fields WH, LD, and RB showed less variability than other fields. In contrast, fields F25, TE, and VN situated in different topographic and agro-ecological regions, $EC_a$ showed high variability. Topographic parameters (*i.e.*, TWI, Slope and Aspect ratio) were extracted from the normalized elevation parameter and varied greatly in LD, RL and VN fields.

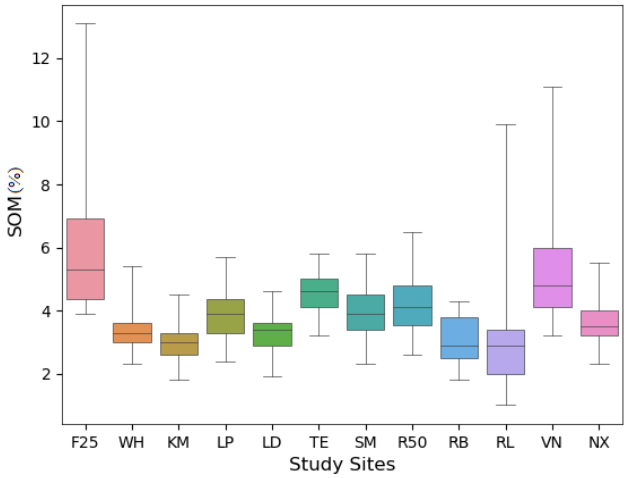
Soil variability was assessed from the results of lab analyses of field samples for pH, BpH, SOM, P, K, and CEC (Figure 4.4). The box plot showed how widely spread the data range is in the measurements (using Min, Max, Median, lower and upper quantiles) and compared the data distributions among the fields. Median values were spread widely for pH, K, and CEC measurements in the twelve fields. Among the 12 agriculture fields, the average pH level was 7.08. BpH measurements in the five fields varied between 6.0 and 7.0. SOM varied between 1% and 13.10% and the most variation was found in F25 and VN fields. The range between maximum and minimum values of P, K, and CEC measurements were also large for the RL field. Standard deviation (STD) for K sample measurements varied greatly from 15.67 ppm to 55.87 ppm. The measured P values in the twelve agricultural fields varied from 6.0 ppm to 134.0 ppm (STD from 3.54 ppm to 24.63 ppm, values were not appeared in Figure 4.4[d]).



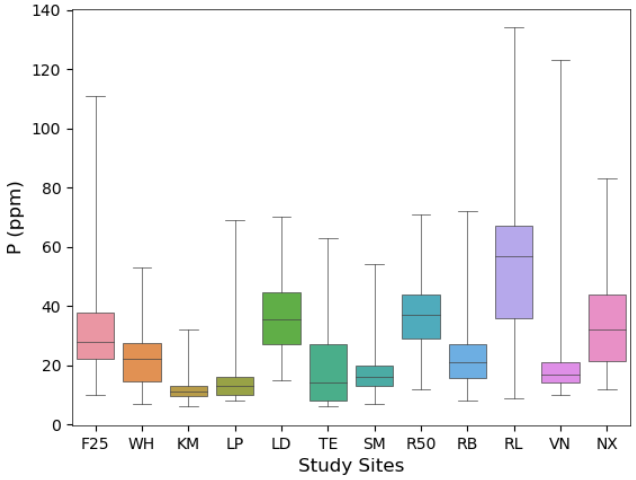

[a] pH in 12 agricultural fields.



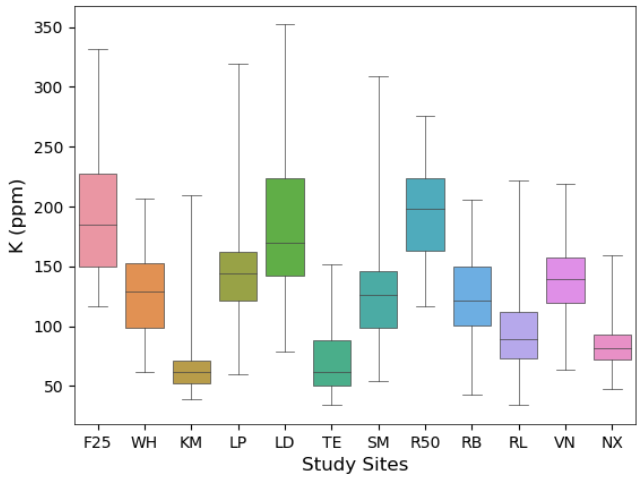

[b] BpH in 5 agricultural fields.



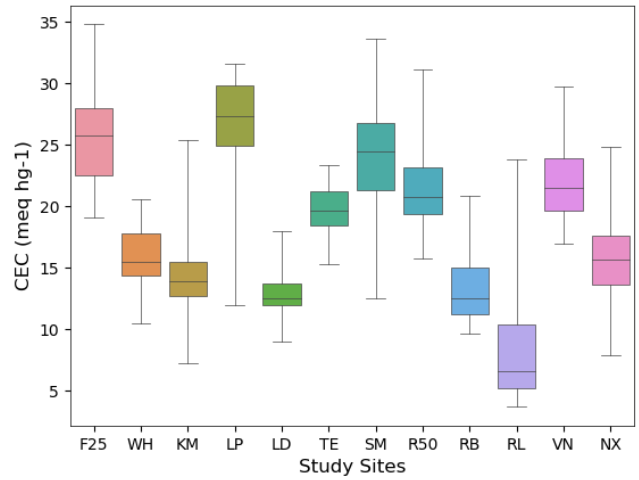

[c] SOM in 12 agricultural fields.




[d] P in 12 agricultural fields




[e] K in 12 agricultural fields.



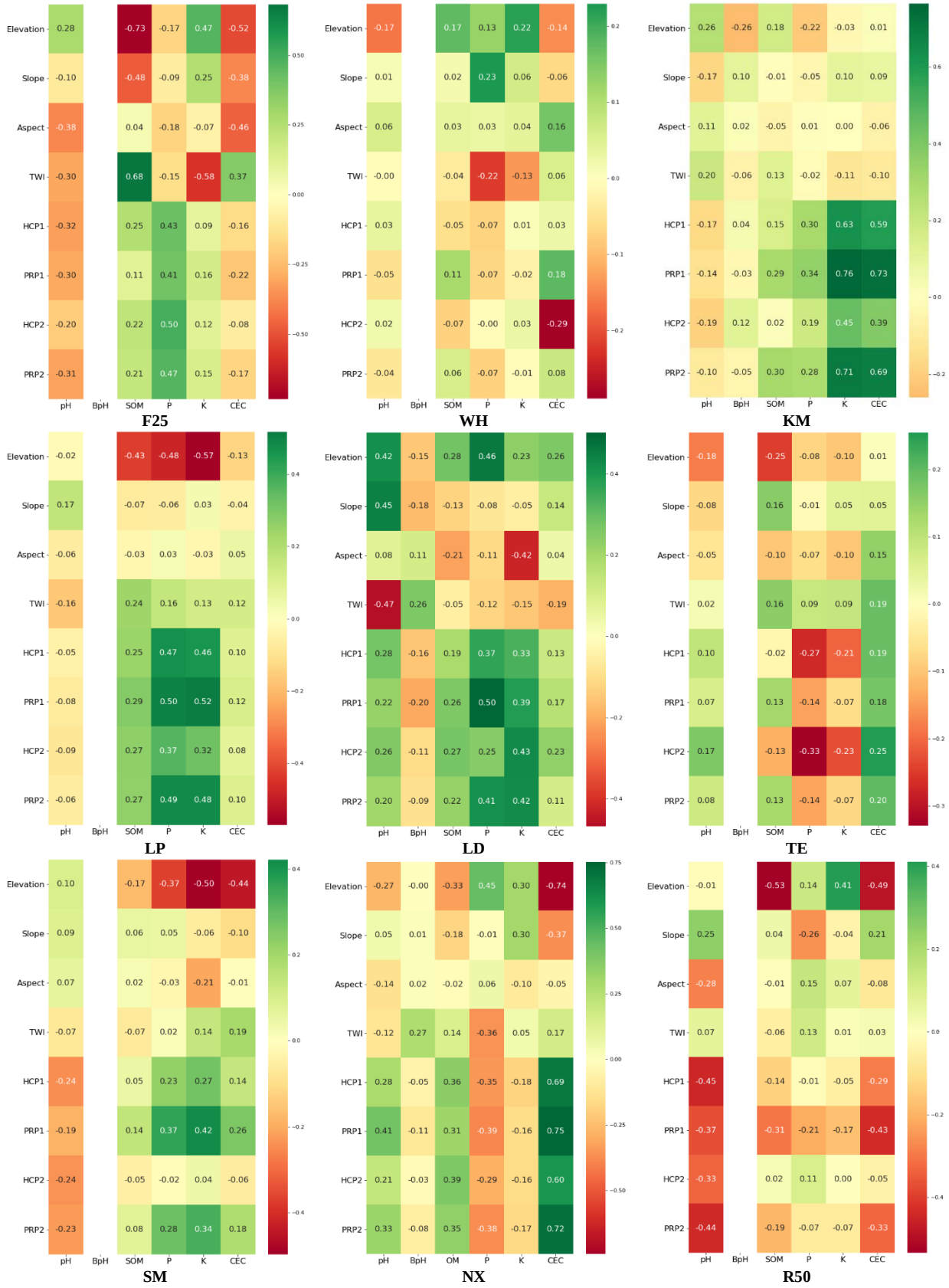

[f] CEC in 12 agricultural fields.

**Figure 4.4** Box plot shows summary statistics for measured soil properties — pH, Buffer pH (BpH), Soil Organic Matter (SOM), Phosphorus (P), Potassium (K), Cation Exchange Capacity (CEC) — in the agricultural fields [a] to [f].

*4.3.2 Correlation analysis and predictive properties*

The level of correlation between sensor variables and soil properties was analyzed to understand any linear relationships existing among the variables (Figure 4.5). The values of BpH were considerably less correlated with the sensor measurements for the five agricultural fields. For the LD field, pH correlated positively with topographic variables (*i.e.*, elevation slope), but no other field had a systematic correlation between $EC_a$ variables and pH. The SOM was negatively correlated with elevation in four fields (*i.e.*, F25, LP, R50, RB), but positively with $EC_a$ in two fields (*i.e.*, RB, RL). Accordingly, SOM can be predicted using the elevation parameters. In two fields (LP and LD) soil phosphorus (P) correlated with shallow (0-1.0 m) $EC_a$ (PRP1 and PRP2) values. P is poorly correlated with $EC_a$ but moderately correlated with topographic parameters. In that case, topographic parameters can also be potentially useful in predicting soil phosphorus. In four agricultural fields (KM, LP, SM, and VN), soil K correlated positively with all $EC_a$ variables, but most strongly ($r$ > 0.70) with shallow (0-0.5 m) $EC_a$ (PRP 1). Therefore, shallow $EC_a$ parameters provided a good predictor for soil K. In four fields (*i.e.*, KM, NX, RL, VN), CEC correlated positively with $EC_a$ variables and showed a particularly strong positive correlation ($r$ = >0.70) with shallow $EC_a$ (PRP1 and PRP2). For fields R50, F25, and NX, the CEC correlated negatively ($r$ = -0.49, -0.52, and -0.74), respectively) with elevation; thus, both topographic and $EC_a$ variables could be useful for CEC prediction.

The predictive capability of the sensor measurements was assessed using regression parameters. The coefficient of determination (Adjusted $R^2$) and standard error of estimate (SE) were reported to the predictive efficiency of the $EC_a$ and topographic auxiliary variables to the various soil properties. Prediction efficiency varied greatly across the twelve study sites. For K prediction, adjusted $R^2$ ($R^2_{adj}$) ranged from 0.01 to 0.64 for all fields. $R^2_{adj}$ was above 0.60 for SOM, K, and CEC (in one field across the Ontario fields). Highest prediction efficiency was found for pH ($R^2$ = 0.54, SE = 0.25 for R50 field), whereas the maximum P prediction value was achieved for RL field ($R^2$ = 0.42). SE of estimates compared to the $R^2$ are discussed in later sections.


Elevation 0.28 -0.73 -0.17 0.47 -0.52
Slope -0.10 -0.48 -0.09 0.25 -0.38
Aspect -0.38 0.04 -0.18 -0.07 -0.46
TWI -0.30 0.68 -0.15 -0.58 0.37
HCP1 -0.32 0.25 0.43 0.09 -0.16
PRP1 -0.30 0.11 0.41 0.16 -0.22
HCP2 -0.20 0.22 0.50 0.12 -0.08
PRP2 -0.31 0.21 0.47 0.15 -0.17
pH BpH SOM P K CEC
F25
Elevation -0.17 0.17 0.13 0.22 -0.14
Slope 0.01 0.02 0.23 0.06 -0.06
Aspect 0.06 0.03 0.03 0.04 0.16
TWI -0.00 -0.04 -0.22 -0.13 0.06
HCP1 0.03 -0.05 -0.07 0.01 0.03
PRP1 -0.05 0.11 -0.07 -0.02 0.18
HCP2 0.02 -0.07 -0.00 0.03 -0.29
PRP2 -0.04 0.06 -0.07 -0.01 0.08
pH BpH SOM P K CEC
WH
Elevation 0.26 -0.26 0.18 -0.22 -0.03 0.01
Slope -0.17 0.10 -0.01 -0.05 0.10 0.09
Aspect 0.11 0.02 -0.05 0.01 0.00 -0.06
TWI 0.20 -0.06 0.13 -0.02 -0.11 -0.10
HCP1 -0.17 0.04 0.15 0.30 0.63 0.59
PRP1 -0.14 -0.03 0.29 0.34 0.76 0.73
HCP2 -0.19 0.12 0.02 0.19 0.45 0.39
PRP2 -0.10 -0.05 0.30 0.28 0.71 0.69
pH BpH SOM P K CEC
KM
Elevation -0.02 -0.43 -0.48 -0.57 -0.13
Slope 0.17 -0.07 -0.06 0.03 -0.04
Aspect -0.06 -0.03 0.03 -0.03 0.05
TWI -0.16 0.24 0.16 0.13 0.12
HCP1 -0.05 0.25 0.47 0.46 0.10
PRP1 -0.08 0.29 0.50 0.52 0.12
HCP2 -0.09 0.27 0.37 0.32 0.08
PRP2 -0.06 0.27 0.49 0.48 0.10
pH BpH SOM P K CEC
LP
Elevation 0.42 -0.15 0.28 0.46 0.23 0.26
Slope 0.45 -0.18 -0.13 -0.08 -0.05 0.14
Aspect 0.08 0.11 -0.21 -0.11 -0.42 0.04
TWI -0.47 0.26 -0.05 -0.12 -0.15 -0.19
HCP1 0.28 -0.16 0.19 0.37 0.33 0.13
PRP1 0.22 -0.20 0.26 0.50 0.39 0.17
HCP2 0.26 -0.11 0.27 0.25 0.43 0.23
PRP2 0.20 -0.09 0.22 0.41 0.42 0.11
pH BpH SOM P K CEC
LD
Elevation -0.18 -0.25 -0.08 -0.10 0.01
Slope -0.08 0.16 -0.01 0.05 0.05
Aspect -0.05 -0.10 -0.07 -0.10 0.15
TWI 0.02 0.16 0.09 0.09 0.19
HCP1 0.10 -0.02 -0.27 -0.21 0.19
PRP1 0.07 0.13 -0.14 -0.07 0.18
HCP2 0.17 -0.13 -0.33 -0.23 0.25
PRP2 0.08 0.13 -0.14 -0.07 0.20
pH BpH SOM P K CEC
TE
Elevation 0.10 -0.17 -0.37 -0.50 -0.44
Slope 0.09 0.06 0.05 -0.06 -0.10
Aspect 0.07 0.02 -0.03 -0.21 -0.01
TWI -0.07 -0.07 0.02 0.14 0.19
HCP1 -0.24 0.05 0.23 0.27 0.14
PRP1 -0.19 0.14 0.37 0.42 0.26
HCP2 -0.24 -0.05 -0.02 0.04 -0.06
PRP2 -0.23 0.08 0.28 0.34 0.18
pH BpH SOM P K CEC
SM
Elevation -0.27 -0.00 -0.33 0.45 0.30 -0.74
Slope 0.05 0.01 -0.18 -0.01 0.30 -0.37
Aspect -0.14 0.02 -0.02 0.06 -0.10 -0.05
TWI -0.12 0.27 0.14 -0.36 0.05 0.17
HCP1 0.28 -0.05 0.36 -0.35 -0.18 0.69
PRP1 0.41 -0.11 0.31 -0.39 -0.16 0.75
HCP2 0.21 -0.03 0.39 -0.29 -0.16 0.60
PRP2 0.33 -0.08 0.35 -0.38 -0.17 0.72
pH BpH OM P K CEC
NX
Elevation -0.01 -0.53 0.14 0.41 -0.49
Slope 0.25 0.04 -0.26 -0.04 0.21
Aspect -0.28 -0.01 0.15 0.07 -0.08
TWI 0.07 -0.06 0.13 0.01 0.03
HCP1 -0.45 -0.14 -0.01 -0.05 -0.29
PRP1 -0.37 -0.31 -0.21 -0.17 -0.43
HCP2 -0.33 0.02 0.11 0.00 -0.05
PRP2 -0.44 -0.19 -0.07 -0.07 -0.33
pH BpH SOM P K CEC
R50

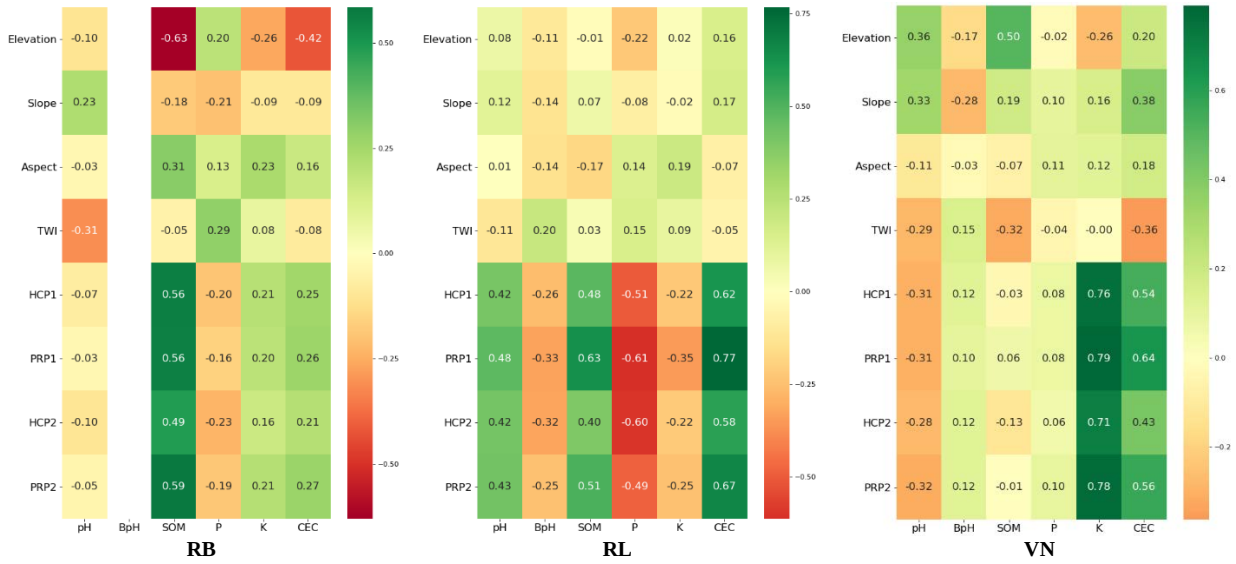
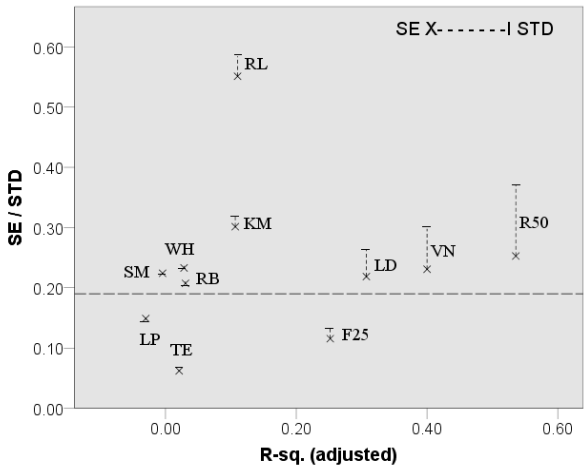


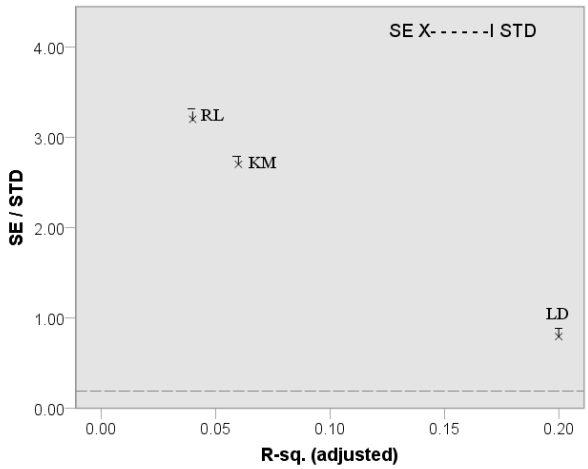

**Figure 4.5** Correlation coefficient I of predictor variables of different soil properties in 12 study sites. The intensity of the green/red color rises with a rise in the negative/positive magnitude of the correlation.

### *4.3.3 Assessment of prediction error for soil properties in Ontario*

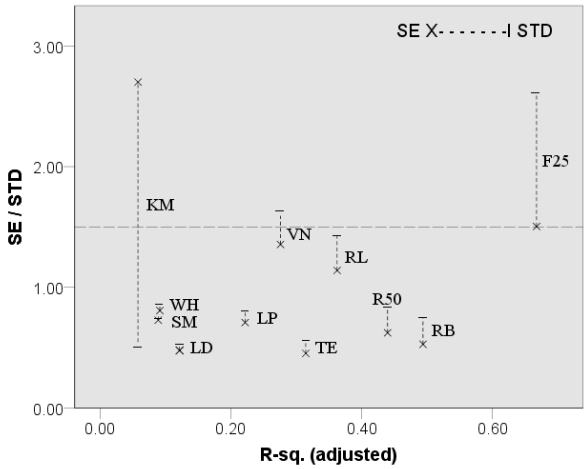
The removal of systematic sensor errors produced by the sensors were minimized in the data preprocessing steps (See section 4.2.2). The following statistical parameters (i.e., STD, SE, and $R^2$) of the samples were reported for the prediction model and error optimization (Figure 4.6). A reference line is added to Figure 4.6 from the median absolute deviation (MAD) of the average reported NAPT results. The reference values of pH and BpH are 0.19. P, K, SOM and CEC values are 18.15 ppm, 29.48 ppm, 1.5% and 5.05 meq $hg^{-1}$, respectively. All the calculated reference values, except pH and K, are higher than the average sensor prediction error in all fields. In this study, the least prediction error (SE) than the STD along with the higher coefficient of determination provides the best prediction model. The less variation (between STD and SE) of the sample estimation that falls under the reference line (MAD of NAPT) could represent very good quality data for building a precise prediction model.



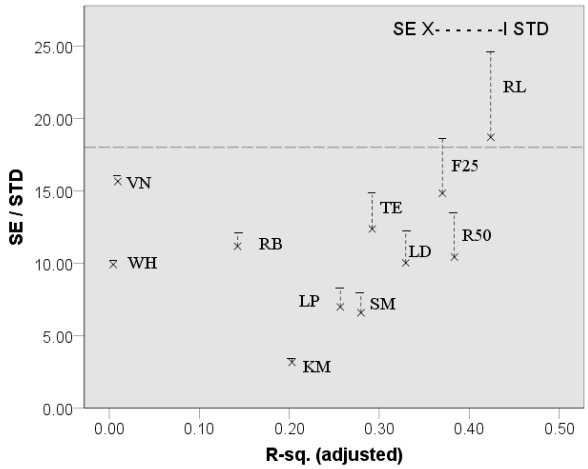

[a] soil pH.



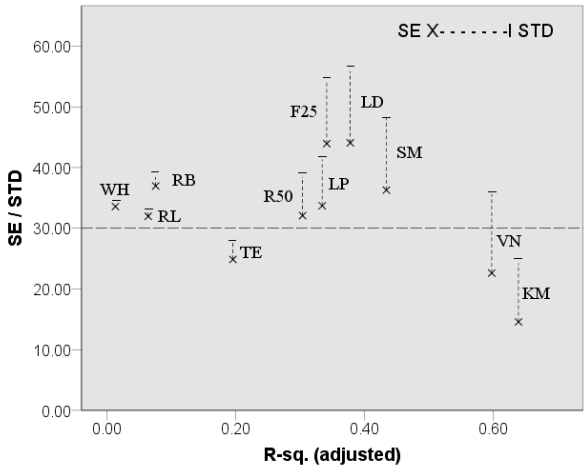

[b] Soil BpH.



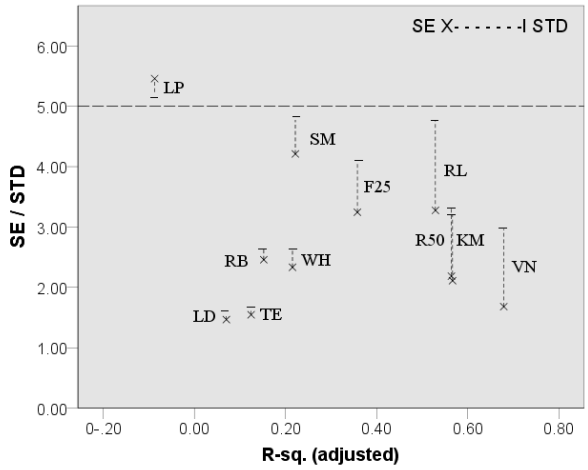

[c] Soil organic matter (SOM).




[d] Phosphorus (P).




[e] Soil potassium (K).




[f] Cation exchange capacity (CEC).

**Figure 4.6** Comparison between the standard error (SE) of estimate and standard deviation (STD) plotted against to the adjusted R-sq.($R^2$) for predicting different soil properties in the 12 agricultural fields [a] to [f].

The $s$ and $s_{\bar{x}}$ of the samples were calculated based on the soil measurements. The prediction error varied according to topographic derivates and DUALEM sensor measurements. In this study, the best prediction model was defined as that generating the least variation between the $s$ and $s_{\bar{x}}$. Model error (difference between $s$ and $s_{\bar{x}}$) along with the $R^2$ adjusted are reported for regional predictions of the targeted six soil properties (Figure 4.6). The data generating less errors (with less variation in estimation) and higher $R^2$ values among the twelve fields are recommended for predicting specific soil properties in Ontario. Lesser errors along with a greater $R^2$ (adjusted) were obtained for the prediction of pH in three fields (*i.e.*, LD, VN, and F25); BpH in the LD field (among five available sites); SOM in four fields (R50, RB, TE, RL); P in five fields (TE, LD, R50, SM, and LP), K in two fields (VN and KM), and CEC in four fields (R50, KM, VN, F25). Among the 12 agricultural fields in Ontario, the VN and R50 fields showed the least errors in soil prediction.

NAPT lab results were published from different laboratory sample analysis across North America and considered as a reference value for all laboratories in Ontario. In Figure 4.6, the average value of the NAPT lab results was considered as a validation line of the soil measurements value for the study sites. When the error estimation ($s$ and $s_{\bar{x}}$) is below the average line of the NAPT values among the study sites, the data can be relied on to make relatively precise predictions, and can be recommended for further soil exploration. With the minimum error consideration, SOM, P, and CEC were predicted often among the Ontario fields.

## 6.5 **Discussion**

The indirect measurements with soil sensors such as RTK GNSS and DUALEM-21S are readily available and provide cost-effective data collection platforms for many provinces in Canada. The data collection environment and our experimental fields were very different in terms of topographic and soil characteristics, and productivity. Preprocessing steps of the high-density sensor data, as described in section 4.2.2, required the development of an optimal prediction model for soil assessment. The descriptive statistics and their analytical results show high-density measurements to be a key element for understanding field variability in terms of soil prediction and mapping. Ji *et al.* (2019) found that the simultaneous measurement of the sensor variables and their large range improved a model's soil prediction capacity. High-density data provided useful

information for making a local, or regional, scale prediction model for Ontario agricultural fields. Topographic and $EC_a$ variables proved useful in predicting several soil properties, including SOM, P, K and CEC. Topographic derivatives along with elevation parameters in some cases provided the data needed for making a quantitative prediction model. Among the topographic variables, only the elevation parameter was suitable for SOM and CEC prediction models. Higher SOM levels were generally found at lower elevations. Soil $EC_a$ measurements, especially for shallow layers (0-1.0 m), were representative of in-field variability and provided useful information for predicting P, K, and CEC. No systematic correlation was found for any sensor variable with respect to pH or BpH. The better prediction capacity associated with sensor measurements could be achieved through the above-mentioned procedures; however, they were maximized while the data collection environments were similar (same temporal or topographic characteristics).

Previous research has shown that lab analysis of large numbers of sample sizes is expensive; nevertheless, it provides a precise assessment of field variability. The present study showed that high-density measurements also provide field variability information along with the optimized samples for making a better prediction model. Among the six soil properties, the overall prediction performance was about 60%. This would reduce the need for the high-density sampling. In the present effort, data optimization using error plotting of several statistical parameters performed better than a single statistical method. Simple correlation and regression techniques were calibrated for high-density sensor measurements, providing better prediction accuracy ($R^2 \geq$ 0.4 for 50% of the fields). The model error varied with topographic and DUALEM sensor measurements as well as among study sites. However, the topographic derivatives combined with the $EC_a$ measurements could assist in constructing a universal prediction model. The model accuracy was compared and validated with the reference value of NAPT results. This comparison protects the model from overfitting and is useful in planning further soil analyses in other study sites. When the sensor prediction and lab measurement errors of the above-mentioned soil properties provided values lower than the NAPT median absolute deviation results, these recommendations could be used for a laboratory certification program in Ontario and in other provinces. 80% of the soil measurement errors (*i.e.*, SOM, P, and CEC) of the fields were below the NAPT results. Among the agricultural fields in Ontario, data from two sites (VN and R50) showed the least errors in predicting all six soil properties.

In this study, factory calibrated DUALEM-21S sensors were used for $EC_a$ measurements in Ontario fields. Huang *et al.* (2018) assessed different PSS instruments for predicting physical and chemical properties. They found that each option could accurately delineate differences in soil physical properties but provided less accurate predictions with regards to phosphorus and potassium content at their site. For the prediction model, we also used a wide range of laboratory soil analysis results from across Ontario, where they followed OMAFRA accredited soil test methods. Our results show SOM, P, and CEC are highly predictable using sensor measurements in the twelve Ontario fields. Among the six properties, SOM and CEC are predicted predominantly using selected sensor measurements (mainly elevation and shallow $EC_a$). Topographic parameters provide promising results for predicting SOM and CEC for some fields. Other topographic parameters can be used for validation purposes in other fields where elevation ranges are similar in the same agro-ecological regions. Overall, the high-density shallow $EC_a$ measurements were key to understanding field variability and represented a substantial input to the prediction model. More BpH data available for the remaining fields and their accurate measurements would enhance the error analysis and model development process in the future. The present research outcomes also suggest strategies to integrate densely measured proximal soil sensing data with the results of laboratory analysis of optimized soil samples, and other data resources. This research highlights the need to develop new sensing technologies and deployment strategies to further increase the accuracy of high-resolution thematic soil maps. Also, NSA field variability maps can be incorporated for minimal soil sampling strategies. Then, the predictive results would be validated in independent study sites and the methods could be assessed to accurately determine how to compensate for accuracy.

## 4.5 Conclusions

This study optimized the modeling process by assessing proximal sensor (high-density apparent electrical conductivity and topographic sensors) data and their prediction capability for the determination of soil nutrients. Even though the study sites have vast elevation differences, topographic derivatives provide promising results for predicting soil organic matter and CEC in Ontario agricultural field soils. This is a general trend for Ontario fields. Shallow $EC_a$ also plays an important role in understanding within-field variability. However, evidence of the applicability

of tested proximal sensing technologies to address spatial variability of certain soil nutrients, such as K, proved to be rather limited.

High-density PSS data plays an important role in soil assessment. Our findings indicate the powerful potential of proximal soil sensing technologies to define the site-specific crop production environments in terms of terrain and soil physical characteristics. The results of the present study suggest that sensor data fusion for multiple soil measurements would be useful in optimizing soil characterization and for improving soil thematic maps. The integration of conceptually different sensors would allow for improved prediction of certain soil properties when compared to a single type of measurement. This continuing research effort will explore additional measurement capabilities that have not been released commercially that could potentially expand the applicability of future proximal soil sensing tools.

Further research will validate and implement results through a set of case studies after which the findings will be disseminated among the agricultural farming communities. The integration of remote sensing and proximal soil sensing techniques could be beneficial to further develop prediction models and thematic maps. This study determined that the protocol of model optimization may be used by commercial sensor users and agronomic service providers to improve their data handling processes and to maximize the information value of the data they generate for their customers. These soil variability and zonal maps can be used to implement variable rate nitrogen fertilization, seeding density, organic fertilizer applications, or liming, thereby, optimizing the use of agricultural inputs by crop producers, their consultants, and agribusiness representatives. A scaled-up adoption of proximal soil sensing technologies would provide advances in agriculture crop production and sustainable natural resource management.

**Funding**

Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) New Directions Research Program (ND2014-2487) and through the Graduate Merit Scholarship, Nature and Technology-FRQNT (B2X), Government of Quebec, Canada.

**Acknowledgements**

W. Ji, H.H. Huang, and S. Lauzon are thanked for their data retrieval and diagnostic work. This research project was also supported by the Grain Farms of Ontario “Precision Agriculture Advancement for Ontario” Project with funding from Growing Forward 2 through Agricultural Adaptation Council. https://gfo.ca/research-projects/c2014ag22.

## Connecting Text to Chapter 5

Chapter 5 is related to the third objective of this study as listed in Chapter 1. Furthermore, there was the indication in Chapter 2 that there is a need for research on the use of sensor-fusion to quantify field-scale soil nutrients. Previous chapters have shown that PSS sensor-based precise soil property prediction requires validation with standard lab-measured values (by standard accredited methods). For that, an integrated PSS platform was used as an example to demonstrate those prediction strategies. In chapter 5, the sensor data combined from different platforms was further optimized in a modeling technique for improvement of prediction quality. A decision-tree model was built with optimized parameters to improve the prediction efficiency of various soil properties at the field scale. The model performance was evaluated by regression prediction results and by observed parameters at different stages.

Initial outcomes were reported and published in the non-refereed conference proceedings listed below and the results were prepared as manuscripts for journal publication:

1. Saifuzzaman, M., Adamchuk, V., Huang, H., & Biswas, A. (2018). Integration of Proximal Soil Sensing and Remote Sensing Data in Agriculture. In *Abstracts from the 39th Canadian Symposium on Remote Sensing 2018, University of Saskatchewan, 19 – 21 June 2018.* Saskatoon, Canada: (Published on-line at https://crss-sct.ca/conferences/csrs-2018).
2. Saifuzzaman, M., Adamchuk, V., Biswas, A., Prasher, S., & Rabe, N., (2019). Geospatial Data Modelling by Integrating Sensor-Fused Data in Agricultural Field Management. In *Proceedings of the 13th Pedometrics conference 'Pedometrics 2019, June 2 – 6.* Guelph ON, Canada: (Published on-line at http://www.pedometrics2019.com).
3. Saifuzzaman, M., Adamchuk, V., & Rabe N. (2020). Sensor-Fusion by Machine Learning Methods for Field-Scale Thematic Soil Mapping. In Abstracts from the 41st Canadian Symposium on Remote Sensing 2020, University of Lethbridge, 13 - 16 July 2020. Lethbridge AB, Canada: (Published on-line at https://crss-sct.ca/conferences/csrs-2020).
4. Saifuzzaman M., Adamchuk, V., and Biswas, A. (2020). Optimization of Random Forest Model for Sensor Data Fusion and Thematic Soil Mapping at Field Scale. *Remote Sensing-MDPI* (In preparation).
5. Saifuzzaman M., Adamchuk, V., Biswas, A., and Prasher, S. (2020). Remote sensing and Proximal Soil Sensor Data Fusion using Geospatial Model for Mapping Agricultural Fields. *Remote Sensing of Environment- Elsevier* (In preparation).

# Chapter 5: Sensor-Fusion through Machine Learning for Field-Scale Thematic Soil Mapping


## Abstract

Sensor-based soil characterization is vital for field management and precision farming practices. To predict and make decisions based on thematic soil properties, agricultural scientists often use high-density proximal soil sensing (PSS) and remote sensing (RS) data. Along with sensor datasets, a subset of soil sampling data can be used to predict soil nutrients in an agricultural field. Accordingly, the present research was designed to develop a prediction framework for sensor-fused data analysis. The potential of integrating proximal soil sensing data with remote sensing imagery to describe field heterogeneity and produce thematic maps with the potential to impose differentiated management decisions was explored. A decision tree-based model was applied to determine soil variability for site-specific crop management. An agricultural field in southern Ontario was selected and mapped using both remote sensing and PSS sensors. The Real-Time Kinematic (RTK) elevation, topographic indices, apparent soil electrical conductivity ($EC_a$), and gamma radionuclide variables were processed, and the data structure evaluated based on summary statistics. RapidEye (Planet Labs, San Francisco, CA, USA) satellite data, visible (VIS)/near-infrared (NIR)/Red-edge (RE) spectrum at a spatial resolution of 5 m, were analyzed to generate vegetation indices used in predictive models. Due to the need to minimize the missing values and adjust discrete data points, a kriging method was used to develop topographic, $EC_a$, and gamma-ray spectrometry-derived maps. To understand soil variability across the field, georeferenced soil samples were collected and used to validate the model. Spectral vegetation indices and other environmental variables derived from PSS and RS data served as model inputs. Descriptive statistical analysis and correlation between sensor variables along with soil sample data enhanced our understanding of spatial heterogeneity.

A random forest regression model with less user influence was designed, and the algorithms were developed in an open-source platform. Model parameters were developed to determine the number of important variables (mostly gamma and topographic variables) and regression trees (optimum number of trees were between 50 and 150) in the training phase that led to optimal model performance and scenario maps. A cross-validation score allowed the evaluation

of the training dataset and improved the predictive accuracy. The coefficient of determination ($R^2$) was about 0.80 and explained maximum variability for predicting pH, K, and Zn. Higher relative prediction errors were reported for SOM, Mn and Ca ($R^2$ = 0.55). Soil nutrient variability determined using sensor-fused data and regression techniques could assist in constructing precise prediction models for soil properties. This research may lead to the development of more accurate thematic soil maps which could improve future site-specific precision farm management.

**Keywords:** Data integration; Geostatistical methods; Random Forest regression; Digital soil modeling, Soil variability map.

## 5.1 Introduction

Proximal soil sensors (gamma-ray spectrometry, apparent electrical conductivity, soil spectroscopy, and yield monitors) and remote sensing sensors (high and low attitude) provide information which facilitates digital soil mapping (DSM) and the characterization of soil ecosystems at various scales (Adamchuk and Tremblay, 2017; Baldoncini *et al.*, 2019; Grunwald *et al.*, 2015; Rouze *el al.*, 2017). Given their individual capability to measure a wide variety of soil profile responses and determine agronomic properties at different scales, remote sensing (RS) and proximal soil sensor (PSS) systems are combined to contribute to site-specific crop and soil management (Adamchuk and Viscarra Rossel, 2011; Grunwald *et al.*, 2015; Söderström *et al.*, 2016). Individual sensors are known to have their limitations and yet, their combined contributions of environmental variables have been increasingly exploited to garner a precise understanding of spatial and temporal heterogeneity (Rizzo *et al.*, 2016). The density of information they provide allows one to document fine-scale soil heterogeneity, which varies at different spatial scales due to several agro-climatic and anthropogenic factors. An understanding of soil variability developed from high-density sensor measurements along with spatio-temporal components, allows a precise determination of physical, chemical and biological soil properties (Hengl *et al.*, 2018). Thus, the precise high-density soil maps of the crop growing environments developed from these data represent a key component in local-scale management decisions.

High-density data integration, or sensor fusion, often incorporates multiple variables to handle the soil environment's spatial and seasonal variations, and to solve agricultural problems. Multiple sources of PSS measurements and their combinations can provide information which allows for the quantification of soil properties and affords a better understanding of an agricultural field. In such efforts, there are several sensor-fusion taxonomies adopted in the DSM and decision-making process (McBratney *et al.*, 2003). Previous studies found that multiplatform data integration outperforms single and integrated multiple variables (Castrignano *et al.*, 2017; Meier *et al.*, 2018). Also, proximal soil sensing (PSS) and remote sensing (RS) sensor fusion provide regional or large field-scale variability while PSS provides only field-scale variability in a DSM (Poppiel *et al.*, 2019). In this integration effort, Grunwald *et al.* (2015) found field sensor data fused with remote sensing indices to correlate with lab-measured values for soil toxicity. Moreover, they found that vegetation indices integrated with $EC_a$ variables could be used to

delineate management zones, whereas other studies built a taxonomic classification using remote VIS-NIR combined with single platform proximal sensor data (Grunwald *et al.*, 2015). To obtain precise results, lab-measured soil information combined with multispectral remote sensing responses is garnering increasing attention in making field-scale soil nutrient predictions (Mulder *et al.*, 2011).

Due to constant changes in the quantity and nature of soil nutrients in an agricultural field, precise soil and vegetation mapping using real, or near-real, time sensor data with multiple environmental variables presents quite a challenge (Brown, 2006; Mahmood *et al.*, 2012; Samuel-Rosa *et al.*, 2015). In the past decade, multivariate statistical modeling was widely applied to evaluate single sensor variables and their relationships with the target properties (Malone *et al.*, 2016; Wadoux, 2019). Later, data fusion processes were employed to synchronize different parameters at various scales and to handle their multiscale uncertainties in geographical space (Grunwald, 2009; Heung *et al.*, 2016; Ji *et al.*, 2019). As a result, data mining algorithms coupled with models were adopted for high-density data processing and for making relatively accurate maps in agricultural research (Padarian *et al.*, 2019; Rasaei and Bogaert, 2019). In the last decade, different prediction frameworks have been proposed for sensor fused data analysis. The data fusion model opens the possibility of integrating geostatistical models to handle many environmental variables along with geospatial data analysis tools (Hengl *et al.*, 2004; Grunwald, 2009; Piikki *et al.*, 2013, Grunwald *et al.*, 2015). Mulla (2013) and Veum *et al.* (2017) proposed advanced sensor fusion algorithms and model optimization for predicting soil nutrients and mapping fertility status at the local level. Previous studies showed that an accurate map enhanced robust decision-making and optimized temporal nitrogen management, organic matter amendments, and the management of other topsoil properties for crop production (Grunwald *et al.*, 2015). Accordingly, a model-based analysis is proposed for accurate soil mapping which, in turn, leads to faster decision-making processes.

A wide variety of fusion approaches have been applied to the assessment of field variability, feature classes, and prediction (Grunwald 2009; Castrignanò *et al.*, 2017). Hierarchical data models are employed to delineate different geospatial variables and map soil classes at the field scale (Sommer *et al.*, 2003). Likewise, supervised learning algorithms, along with classification and regression tree (CART) approaches, represent powerful supervised learning

methods that are widely used, in bioinformatics and many other fields for multivariate data analysis and faster decision-making (Breiman, 2001; Qi, 2012). In addition to their application in medical and remote sensing analysis, random decision forests and regression tree models (Minasny and McBratney, 2016; Witten *et al.*, 2017) are increasingly drawing attention for assessing many variables and their multidimensional relationships in agricultural research. Tree-based models were applied for sensor fusion and classification purposes in several DSM applications (Heung *et al.*, 2014; Grunwald *et al.*, 2015; Brogi *et al.*, 2019), wherein the algorithm evaluated errors produced in different training stages and predicted model efficiency through residuals modeling (Wadoux, 2019; Pouladi *et al.*, 2019). Such models can handle unbalanced/missing datasets, are more stable, have faster runtimes, and provide robust data in weighing classified samples iteratively in remote sensing data classification (Mulder *et al.*, 2011; Pelletier *et al.*, 2016). In this study, a tree-based regression model handled sensor-fused data and assessed their complex relationship in support of a DSM effort.

In machine learning models, preparing training data from various sensors and the optimization of model parameters (hyperparameters) are key tasks in achieving accurate decision making and predictions (Grimm *et al.*, 2008; Guo *et al.*, 2015; Keskin *et al.*, 2019). A geostatistical analysis is applied to standardize various sensor variables and determine the training dataset (Szatmári and Pásztor, 2019). The regression tree model controls the selection of variables from the training samples and is efficient in handling errors (Blanco *et al.*, 2018; Pelletier *et al.*, 2016). Hengl *et al.* (2004) and Heung *et al.* (2016) applied a classification model with many training datasets to predict a wide range of soil properties on a regional scale. Likewise, others analyzed many environmental covariates and then they were used as input training samples to attain the best prediction results (Vermeulen and Niekerk, 2017; Zeraatpisheh *et al.*, 2019). In the training phase of the prediction framework, different optimization techniques have been used to estimate model parameters and reduce prediction uncertainties (Xiong *et al.*, 2015; Dharumarajan *et al.*, 2017; Merrill *et al.*, 2017; Vaysse and Lagacherie, 2017). While such methods have been adopted for regional prediction from a large-scale dataset (Rad *et al.*, 2014; Minasny and McBratney, 2016), there remains a need to implement a regression tree model for local or farm-scale applications. The present research assesses training datasets and model parameters to generate scenario maps and predict soil properties at a local scale. Accurate model-estimated soil properties could help

optimize the use of agricultural inputs and make farms more profitable and sustainable by decreasing water and fertilizer consumption.

The goal for this research was to evaluate a supervised learning algorithm that integrates PSS and RS sensor data along with field measurements and assesses their hierarchical relationship for digital soil mapping. In this effort, a random forest model was applied to combining field surface and subsurface measurements to determine soil variability at the field scale. The model was also assessed with respect to the regression parameters of the observed variables at different stages and to determine its behavior in digital soil mapping. The specific objective of this research was to develop a prediction framework for sensor fused data analysis and modeling. Modeling explores the potential of integrating proximal soil sensing data with remote sensing imagery to delineate field heterogeneity and produce thematic maps suitable for potentially differentiated management decisions. A better understanding of field heterogeneity in a landscape and the production of accurate soil maps helps farmers and other land managers to optimize their decision-making process and to develop profitable and sustainable environmentally friendly operations.

## 5.2 Materials and methods

### *5.2.1 Experimental site*

A 39.5 ha agricultural field, situated at the Woodrill Farms near Guelph, Ontario, Canada was selected and mapped using both remote sensing and PSS sensors (Figure 5.1). The soil texture was mainly loam, which maintains very good drainage conditions. According to the Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA) database, soybean [*Glycine max* (L.) Merr.] and corn (*Zea mays* L.) are the targeted annual crops in the region. Figure 5.2 shows the methodological development of this research.

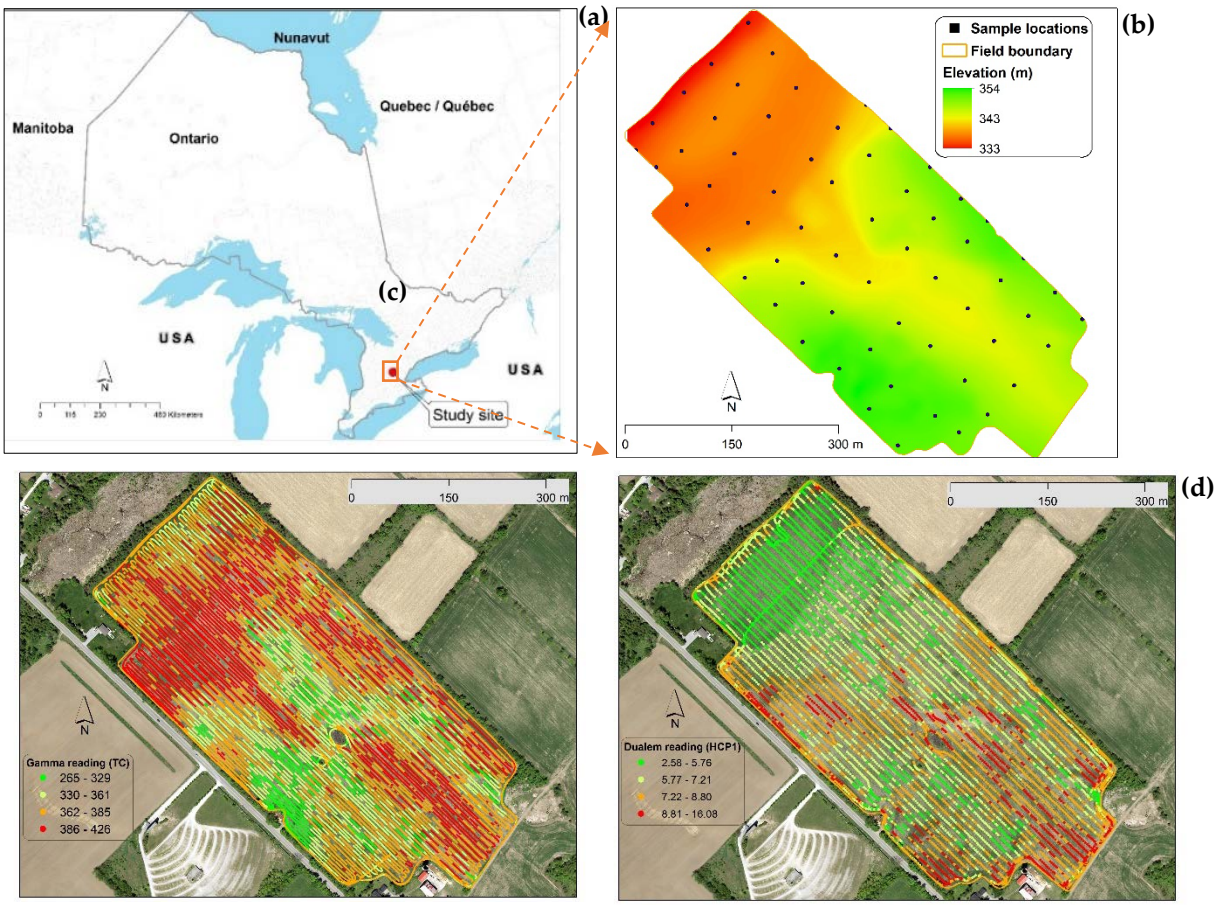



**Figure 5.1** (**a**) Location of study site in Ontario, Canada, **(b)** terrain model along with soil sample locations at the study site, and field boundary with sensor measurements (aerial image on the background): (**c**) gamma-ray sensor reading, and (**d**) soil apparent electrical conductivity ($EC_a$).

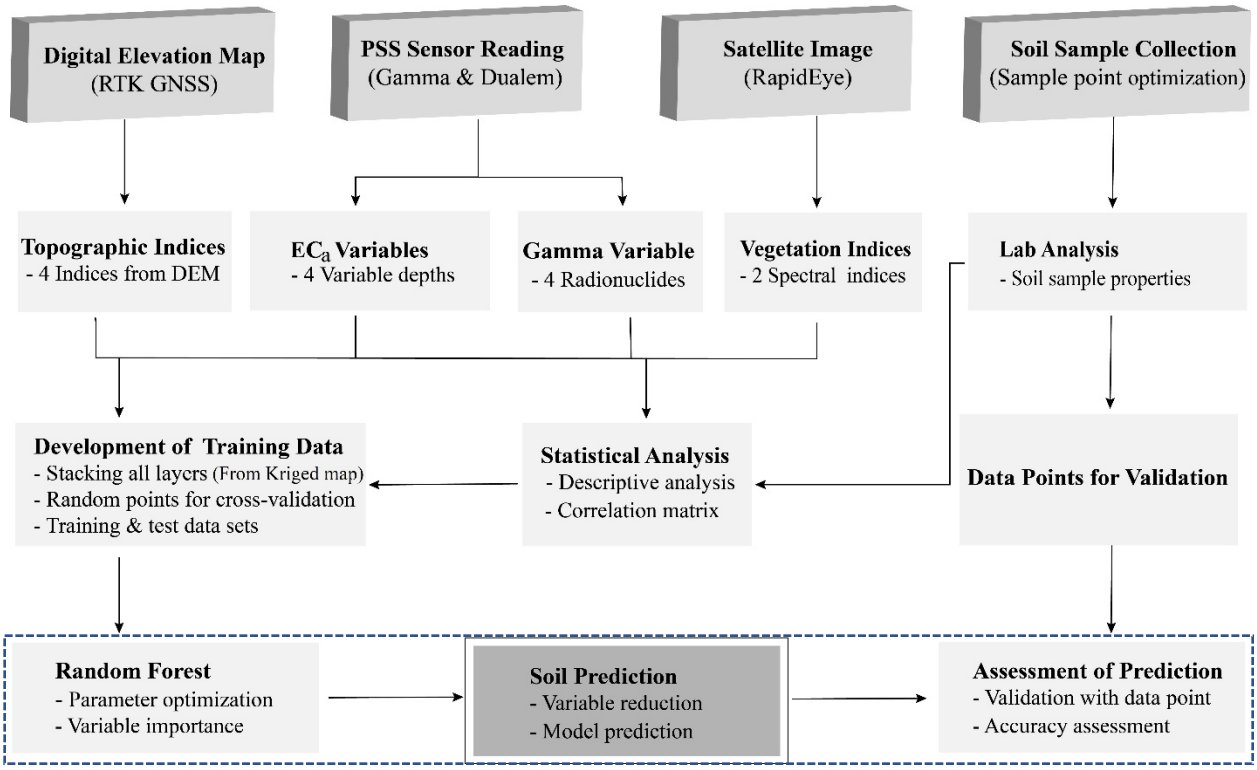



**Figure 5.2** Flowchart showing methodological development (*i.e.*, data collection, processing, training data sets and soil prediction model and accuracy assessment) in this research. The model development parts (dotted line) were described in detail in later section.

### *5.2.2 Soil sensing by proximal soil sensors*

The study site was mapped using DUALEM-21S together with a Real-Time kinematic (RTK) global navigation satellite systems (GNSS) receiver, and a gamma-ray spectrometer, thereby, generating high-density field variability maps. A vehicle equipped with two types of proximal soil sensors (RTK GNSS and DUALEM-21S) was used for topographic and apparent soil electrical conductivity ($EC_a$) mapping in August 2015. The measurements were recorded every 0.1 s with a vehicle travel speed around 10 km $h^{-1}$. The measured data points of elevation and $EC_a$ were at intra- and inter-row spacing of approximately 5 m and 10 m, respectively. Despite diverse data sources and various data standardization among the industries, generic rules were developed in terms of data format and preprocessing steps to assess the PSS data sensitivity to bare soil properties. Timestamps, locations, speed of the sensor vehicle, the distance between data points, and other variable measurements were evaluated in the preprocessing steps. Detailed procedures were discussed in Ji *et al.* (2017). Potential outliers and null values of the PSS

measurements were identified in this step, and about 8% of the PSS data was removed. In this study, different environmental variables were considered for building the input and training datasets used by the model. General statistical analysis and correlation matrices of the selected variables are used to determine targeted variables in the following sections.

#### 5.2.2.1 Soil sensing – $EC_a$ measurements

The $EC_a$ data ($n$ = 6,931 points) were obtained using an electromagnetic induction instrument (DUALEM-21S, Dualem, Inc., Milton, ON, Canada). The instrument (with two-pairs of electromagnetic receivers: horizontal co-planar geometry-HCP and perpendicular geometry-PRP) served to collect the soil apparent electrical conductivity ($EC_a$) of four different depths: HCP1 – 0-1.6 m ($EC_a^{0\cdot1.6}$), PRP1 – 0-0.5 m ($EC_a^{0\cdot0.5}$), HCP2 – 0-3.2 m ($EC_a^{0\cdot3.2}$), and PRP2 – 0-1.0 m ($EC_a^{0\cdot1.0}$). Descriptive statistics [Minimum (Min), median, Maximum (Max), Standard deviation (STD), and mean] were generated from the measurements for sensor data assessment (Table 5.1). Values of $EC_a$ for different soil depths – $0.72 \leq EC_a^{0\cdot0.5} \leq 14.12$, $1.08 \leq EC_a^{0\cdot1.0} \leq 14.60$, $2.58 \leq EC_a^{0\cdot1.6} \leq 16.08$, and $0.50 \leq EC_a^{0\cdot3.2} \leq 14.44$ mS m$^{-1}$ – were determined by statistical analysis and reflected field variability at the small site.

**Table 5.1** Descriptive statistics of four DUALEM-21S sensor readings: $EC_a^{0\cdot1.6}$, $EC_a^{0\cdot0.5}$, $EC_a^{0\cdot3.2}$, and $EC_a^{0\cdot1.0}$ mS m$^{-1}$.

| Sensor measurements | Min | Median | Max | Mean | STD |
|---|---|---|---|---|---|
| $EC_a^{0\cdot1.6}$ | 2.58 | 6.90 | 16.08 | 6.96 | 1.55 |
| $EC_a^{0\cdot0.5}$ | 0.72 | 4.44 | 14.12 | 4.55 | 1.38 |
| $EC_a^{0\cdot3.2}$ | 0.50 | 4.44 | 14.44 | 4.61 | 1.85 |
| $EC_a^{0\cdot1.0}$ | 1.08 | 4.68 | 14.60 | 4.75 | 1.50 |

#### 5.2.2.2 Soil survey – Topographic mapping and derivatives

Topographic data ($n$ = 7,110 points) were collected from the agricultural field using a Trimble AgGPS 542 GNSS receiver and base station (Trimble, Inc., Sunnyvale, California, USA). Topographic variations were determined by statistical analysis. The field elevations ranged from 333 to 354 m with a standard deviation of 5.76 m. Slope and aspect ratio (AR) ranges were derived from the maximum elevation, while the topographic wetness index (TWI) was derived from a

digital elevation model (DEM) of the study site. Besides ArcGIS v10.7 (ESRI, Redlands, California, USA), a commercial software package used in geospatial analysis of topographic variables, SAGA GIS v6.3.0 (Department for Physical Geography, Hamburg and Göttingen, Germany), an open-source software tool, was used for calculating TWI. TWI and AR were calculated as follows:

$$TWI = \ln \frac{a}{\tan \beta} \quad (1)$$

where, $a$ is the upland contributing area, [(flow accumulation $+ 1$) $\times$ cell size], and $\beta$ is the slope in radians.

$$AR = \sin \frac{aspect}{2} \quad (2)$$

where *aspect* is derived from maximum/minimum elevations.

5.2.2.3 Soil survey – gamma-ray sensing

The study site was also mapped with a gamma-ray (γ−ray) sensor (SoilOptix®, Tavistock, ON, Canada). At 60 cm above the soil surface, the sensor was mounted on a vehicle and collected points continuously, following parallel lines 12 m apart. The data was logged every second and the measurements were continuously recorded with a travel speed of 10 km $h^{-1}$; 26,080 data points were collected ($n$ = 20,129 were used after preprocessing of the data) in July 2015. This non-invasive sensor measured four γ−ray spectra (radionuclides) [Uranium-238 ($^{238}$U), Thorium-232 ($^{232}$Th), and Potassium-40 ($^{40}$K), and Total count (TC)] in becquerel per kilogram (Bq $kg^{-1}$) (Dierke and Werban, 2013; Mahmood, *et al.*, 2013). The range between maximum and minimum values of the four radionuclides was very large and used to assess the variability of the field (Table 5.2). Average values of TC, $^{40}$K, $^{238}$U, and $^{232}$Th were 371.31, 354.10, 20.03, and 19.97 Bq $kg^{-1}$ respectively (with a standard deviation of 24.93, 49.63, 4.86, and 3.75 Bq $kg^{-1}$).

**Table 5.2** Descriptive statistics of four measured γ−ray radionuclides (Bq $kg^{-1}$) from the agricultural field in Ontario.

| Sensor measurements | Min | Median | Max | Mean | STD |
|---|---|---|---|---|---|
| TC | 264.73 | 376.00 | 425.64 | 371.31 | 24.93 |
| $^{40}K$ | 142.06 | 356.11 | 515.26 | 354.10 | 49.63 |
| $^{238}U$ | 5.09 | 20.00 | 40.71 | 20.03 | 4.86 |
| $^{232}Th$ | 5.66 | 19.99 | 35.11 | 19.97 | 3.75 |

*5.2.3 Satellite data and derived indices*

A RapidEye satellite image along with two (Orthophoto and Dove) high-resolution datasets were collected to analyze bare soil and vegetation characteristics (Table 5.3). Remote sensing image processing steps were followed (*e.g.*, radiometric correction, stitching, co-registration, stack bands, etc.). In this study, orthophoto and Dove images, with three visible multispectral bands, were used only for co-registration of multiple RapidEye images and for assessing derived vegetation indices. Also, the bare soil orthophoto was used for identifying field sampling locations. In addition to the traditional visible (RGB) and near-infrared (NIR) spectral bands, RapidEye imagery presented a red edge part of the spectrum as well. The two popular and standardized spectral indices, Normalized Difference Red Edge Index (NDRE), and Normalized Difference Vegetation Index (NDVI), were derived from RapidEye satellite data to identify the strong absorption spectrum of chlorophyll and defined as:

$$\mathrm{NDVI} = \frac{\rho NIR - \rho Red}{\rho NIR + \rho Red} \tag{3}$$

$$\mathrm{NDRE} = \frac{\rho NIR - \rho RedEdge}{\rho NIR + \rho RedEdge} \tag{4}$$

where, Near-infrared band ($\rho NIR$), Red band ($\rho Red$), and RedEdge band ($\rho RedEdge$) were used for the index.

**Table 5.3** Remote sensing data characteristics and their sources.

| Satellite Sensor | Pixel (m) | Spectral Bands & Wavelength (nm) | Imaging Date | Source |
|---|---|---|---|---|
| OrthoPhoto | 0.2 | - | 23 May 2015 | OMAFRA/OMNRF[1] |
| Dove | 3.0 | Blue: 420 – 530<br>Green: 500 – 590<br>Red: 610 – 700<br>NIR: 770 – 900 | 30 July 2017 | Planet Labs[2] |
| RapidEye | 5.0 | Blue: 440 – 510<br>Green: 520 – 590<br>Red: 630 – 685<br>Red Edge: 690 – 730<br>NIR: 760 – 85 | 09 August 2017 | Planet Labs[2] |

[1]Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) and Ontario Ministry of Natural Resources and Forestry (OMNRF).

[2] Planet Labs, Inc. in San Francisco, USA (https://www.planet.com)

### *5.2.4 Spatial interpolation and point data extraction*

Interpolation using spatial autocorrelation was performed to understand the variability of point-based PSS data and the spatial characteristics of the missing values. Ordinary Kriging interpolation maps were generated from the spherical variogram model and data structure of all sensor measurements in ESRI ArcGIS software (v10.7). Elevation data points were interpolated for making the digital elevation model. Four gamma nuclides and four $EC_a$ data pointed were also interpolated to facilitate the data extraction process. Multiple kriged maps were delivered spatial covariates associated with the sampling points into a data file (text file) as a software requirement to run the RF model. Finally, the text data file containing multiple layers of sensor variables, sensor-derived variables, and soil measurements was used to assess the model parameters and train the prediction model.

### *5.2.5 Soil sampling and laboratory analysis*

Based on the variability of RTK GNSS and DUALEM-21S sensor measurements in the agricultural field, an optimum number of soil samples were collected for the laboratory analysis. In this research, a Zonesmart system-based 1-acre grid sampling strategies was applied. Based on

the maximum field variability derived from the previously collected PSS measurements, the sample location was placed in each grid. A total of 62 targeted sampling points were selected from the grid centers, with an average sampling density of 5 samples per hectare. The center points were then positioned using the orthophoto and a Garmin handheld GPS (wide area augmentation system – WAAS corrected). At each location, the soil samples were collected from a close radius of 6-10 cores with an approximate depth of 15 cm. Soil samples were collected from the site at the beginning of the cropping season (August 2015).

The lab measured soil analysis data were processed and selected for the prediction model (Table 5.4). In this study, the lab-measured, soil micro- and macro-nutrients, were pH, soil organic matter (SOM), extractable phosphorus (P) and potassium (K), cation exchange capacity (CEC), Magnesium (Mg), Manganese (Mn), Zinc (Zn), and Calcium (Ca). Various soil test methods were used for analyzing all field samples: pH – 1:2 saturated paste; OM% - Walkley-Black (0-8%), Loss on Ignition (>8%); P – Olsen sodium bicarb; K/Mg/Ca – ammonium acetate extract; Mn – Phosphoric acid extract; Zn – DTPA extract; and CEC – calculated by converting soil test K/Mg/Ca to milliequivalents.

**Table 5.4** Descriptive statistics of laboratory measured nine soil properties.

| Soil properties from sample analysis | Descriptive statistics | | | | |
|---|---|---|---|---|---|
| | **Min** | **Median** | **Max** | **Mean** | **STD** |
| pH | 6.50 | 7.50 | 7.90 | 7.44 | 0.26 |
| Soil organic matter (SOM) % | 1.90 | 3.40 | 4.60 | 3.38 | 0.51 |
| Soil Phosphorus (P) ppm | 15.00 | 35.50 | 70.00 | 36.90 | 12.24 |
| Soil Potassium (K) ppm | 79.00 | 169.50 | 352.00 | 183.79 | 55.87 |
| Magnesium (Mg) ppm | 123.00 | 271.50 | 393.00 | 267.66 | 63.77 |
| Calcium (Ca) ppm | 1236.00 | 1700.50 | 2995.00 | 1741.89 | 276.01 |
| Zinc (Zn) ppm | 1.50 | 4.50 | 17.00 | 5.98 | 3.69 |
| Manganese (Mn) ppm | 12.60 | 16.70 | 25.00 | 17.19 | 2.73 |
| Cation exchange capacity (CEC) meq $hg^{-1}$ | 9.00 | 12.50 | 18.00 | 12.58 | 1.52 |

Descriptive statistics for lab analysis-derived parameters yielded density estimation plots (Figure 5.3), which showed the variability existing among the soil properties measured at the study site. The range (minimum and maximum values), standard deviation ($\sigma$), and mean ($\mu$) of the data for each soil parameter showed large variability in the data structure. The measured pH value varied between 6.50 and 7.90 with a $\sigma$ of 0.26 and mean ($\mu$) of 7.44. For the SOM measurements,

the high-density and lowest density occurred near the value of 3.5% and 2.5%, respectively, where μ = 3.38. CEC measurements varied between 9 and 18 meq $hg^{-1}$. The range of Mg measurements varied between 123 to 393 ppm. Moderate variability was found in P measurements (μ = 36.90 ppm), while K and Ca showed high variability (ranges of 273 ppm and 1759 ppm, respectively) in the field. The wider range of the sensor response (predictor variables as described in section 5.2.2) was applied to develop the soil prediction models for the agricultural field. Based on the moderate to high variability in the soil nutrients, the predictor variables were used to build the prediction model.

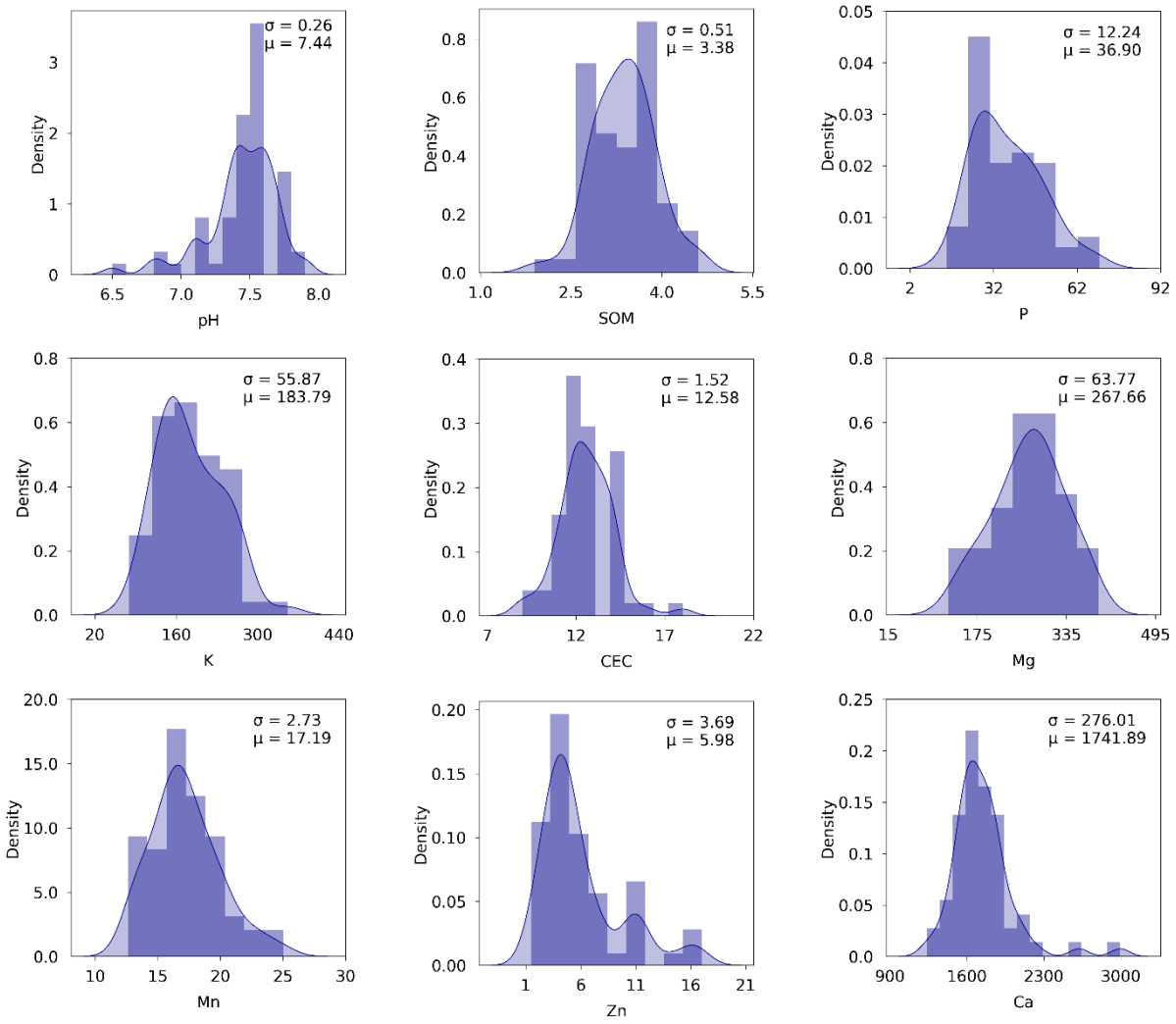



**Figure 5.3** Density plots showing the distribution of soil sample measurements for the field under study.

*5.2.6 Environmental covariates for prediction*

A total of fourteen environmental covariates – topographic indices, soil electrical conductivity, gamma radionuclides concentration, and multispectral vegetation indices – were independently assessed based on the sensor's characteristics (Table 5.5) and prepared as predictor variables for this study. All environmental variables were prepared for a point-based, targeted sampling grid, prediction of each soil property. All high-density sensor data were interpolated and extracted using the sampling points as discussed in the section 5.2.4.

**Table 5.5** The environmental covariate derived from different sensors and prepared as predictor variables.

| **Predictor variables** | **Sensor sources** | **Data captured** |
|---|---|---|
| *Remote sensing attributes*<br>NDVI<br>NDRE | RapidEye satellite | August 2017 |
| *Topographic indices*<br>Elevation (m)<br>Slope %<br>AR – Aspect ratio<br>TWI – Topographic wetness index | RTK GNSS | August 2015 |
| *Gamma-ray (Bq $kg^{-1}$)*<br>TC – Total count<br>$^{40}K$<br>$^{238}U$<br>$^{232}Th$ | Gamma-ray | July 2015 |
| *Soil $EC_a$ (mS $m^{-1}$)*<br>HCP1 (0-1.6 m)<br>HCP2 (0-3.2 m)<br>PRP1 (0-0.5 m)<br>PRP2 (0-1.0 m) | DUALEM-21S | August 2015 |

*5.2.7 Statistical analysis and relationship among the variables*

Multicollinearity assessed the spatial data correlation among the predictor variables (sensor variables and sensor derived variables). High collinearity was found mostly among $EC_a$ variables (Figure 5.4), although these were measured for different depths – HCP1 (0-1.6 m), HCP2 (0-3.2 m), PRP1 (0-0.5 m), and PRP2 (0-1.0 m). The change of $EC_a$ magnitude in variable depths might

be useful for the characterizing of a soil profile. The slope was highly correlated with TWI only, but less so with the remaining topographic variables.

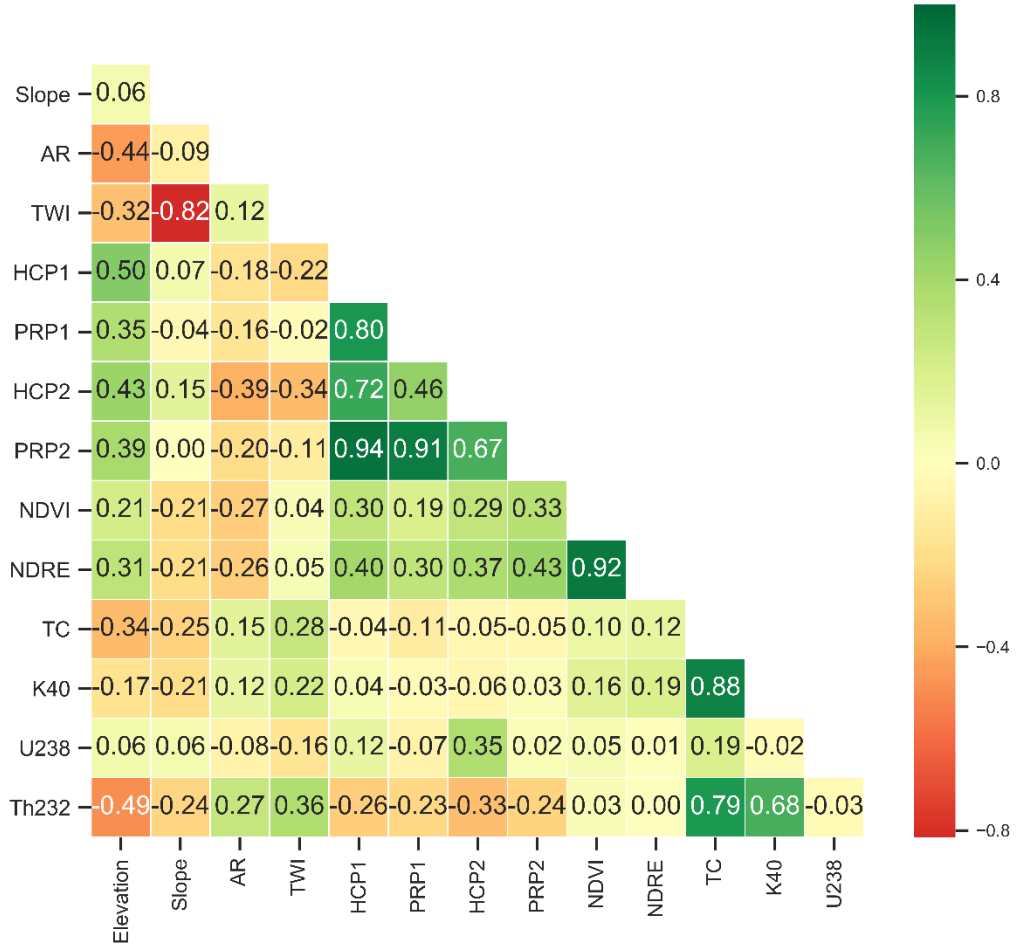



**Figure 5.4** Correlation matrix showing the collinearity among predictor variables. Color intensity increases with higher negative (-) and positive (+) Pearson's correlation values.

Due to the multi-directional linear relationship between several sensor variables and the soil measurements, it is challenging to evaluate the prediction capacities of the sensor measurements for a specific soil property; therefore, a model is needed which is capable of handling hierarchical relationships. Hence, a machine learning regression model, which handles the data fusion process and the complex relationships, was proposed to assess the predictive capacities.

*5.2.8 Modeling techniques and prediction framework*

A supervised machine learning model served as a framework for soil parameter prediction. The supervised method was capable of assessing various complexities at the local scale and it was used for predicting small datasets (Huang *et al.*, 2014). Environment-soil covariates were used in the machine learning prediction framework to predict unknown location values. Each soil nutrient prediction was produced with model validation and accuracy assessment procedures.

5.2.8.1 Random forest (regression tree) prediction

Random forest (RF), a supervised and tree-based ensemble method, was used for soil parameter prediction (Breiman, 2001; Huang *et al.*, 2014). This non-parametric model is easy to understand and requires few user inputs. An RF model is the enhanced version of the regression tree model and its deterministic behavior is assessed here with respect to model fitting and prediction of soil maps (Figure 5.5). The advanced algorithm in the random forest method is better at addressing the large data classification issues and regression, allowing for good estimation of soil parameters with the view of solving agricultural problems (Hengl *et al.*, 2017). This model can handle missing values and large data sets of high dimensionality, while showing high accuracy in mapping and prediction. This model establishes a hierarchical relationship between the multi-sensor variables and the soil nutrients and takes an average of all individual decision tree estimations. Python v3.6 was used, *i.e.*, RandomForestRegressor from the *scikit-learn* package (Géron 2017).

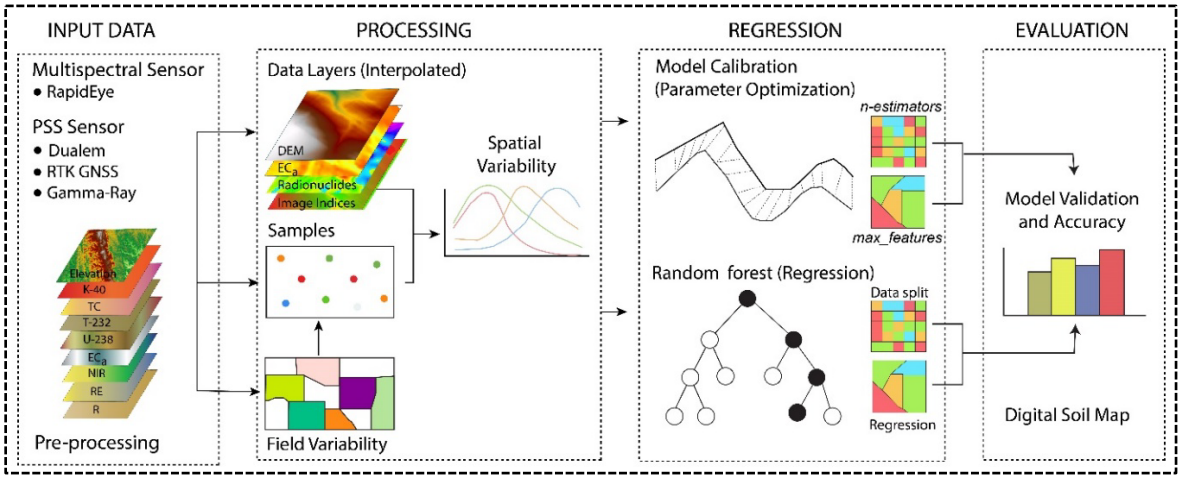



**Figure 5.5** This diagram described random forest model development (partially illustrated in Figure 2) and components: data input and processing, regression and model validation.

In the model training procedure, the bootstrap aggregated approach drew randomly selected samples from the grid samples (with replacement) to build a decision tree. In the training phase, in-bag samples (70% of the training data) served to train the model and out-of-bag (OOB) samples (approximately 30% of the training data) served to do cross-validation (CV). Bootstrap aggregation methods in the forest model resampled training dataset and created node splits for decisions (Hengl *et al.*, 2018). Computational time may vary based on input variables and the number of splitting nodes. The training datasets for this supervised learner originated from variable importance and were used for model building.

For the regression procedure, the random forest-built *k* trees, where the predicted values were the average of all individual tree predictions. However, it does not predict the value which is beyond the training samples. Random forest regression creates a set of *K* trees [$Tx_i$,….., $Tx_k$)], where $x = [x_i,\ldots..,\ x_\beta]$, is a β-dimension of the input vector which forms a forest. The predicted values are obtained by the aggregation of the results of all individual trees. The following equation provides the random forest regression predictor:

$$f(x) = \frac{\sum_{k=1}^{k=K} T_k(x)}{K} \tag{5}$$

Random forest builds a set of regression trees (*K*) and averages the predictions of individual trees to make a final prediction. Where *k* is the individual bootstrap sample and $T_k$ is the individual learner or decision tree.

For a random forest individual tree $T_k(x)$ construction (Zhou *et al.*, 2019; Hengl *et al.*, 2018) the following equation applies:

$$T_k(x) = t(x; t_{xi}, \cdots, t_{xK}) \tag{6}$$

For each number of trees constructed, bootstrap samples (*k*) are drawn for a new training set with a replacement from the original training data set. As a result, a regression tree then builds from the randomized drawn training sample of the original data. The $t_{xK}$ ($k$ = 1, 2, …$K$) is the $k^{th}$ training sample with a pair of values, which produces the target variable ($y$) and covariates ($x$), where $t_{xi} = (x_k , y_k)$. The OOB sample is used for CV (testing). Independent validation using the OOB sample contributes to making a robust forest model.

Based on the environmental variables described in Tables 5.4 and 5.5, a RF trend model was developed. Training data parameters drawn from random sample selection with replacement and their optimization are discussed in the following section.

5.2.8.2 Development of training data and optimization of hyperparameters

Bootstrap sampling and its hyperparameter must be optimized to prevent model overfitting. A random sample selection with replacement within the training set provides two important model-building parameters: (*i*) the number of trees (*n-estimators*) or decision trees grown for the regression predictors, and (*ii*) the number of predictor variables (*max_features*), which are randomly sampled in each binary split and yield the best split from the random subset. As the most important tuning parameter, *max_features* is optimized during the training phase by the user (Heung *et al.*, 2014). The forest tree is grown until the node variance is minimized and then tuned in the training phase (Figure 5.6). Without tuning the parameters, building a lot of trees and splitting notes in the training phase slowed down the computing process. Upon selection of the optimum parameter values for each level the *n-estimators* and *max_features*, model calibration is performed to report the error and model efficiency. Heung *et al.* (2014) show that OOB error in the RF model is a better estimator of error than the CV in optimizing model parameters. Based on preliminary results, *n-estimators was* selected to estimate the stable OOB error rate and determine if it was small enough (*e.g.*, *n-estimators* = 50*)* to increase computation efficiency. By default, *max_features* are chosen for all variables when the model makes the best split in the training phase.

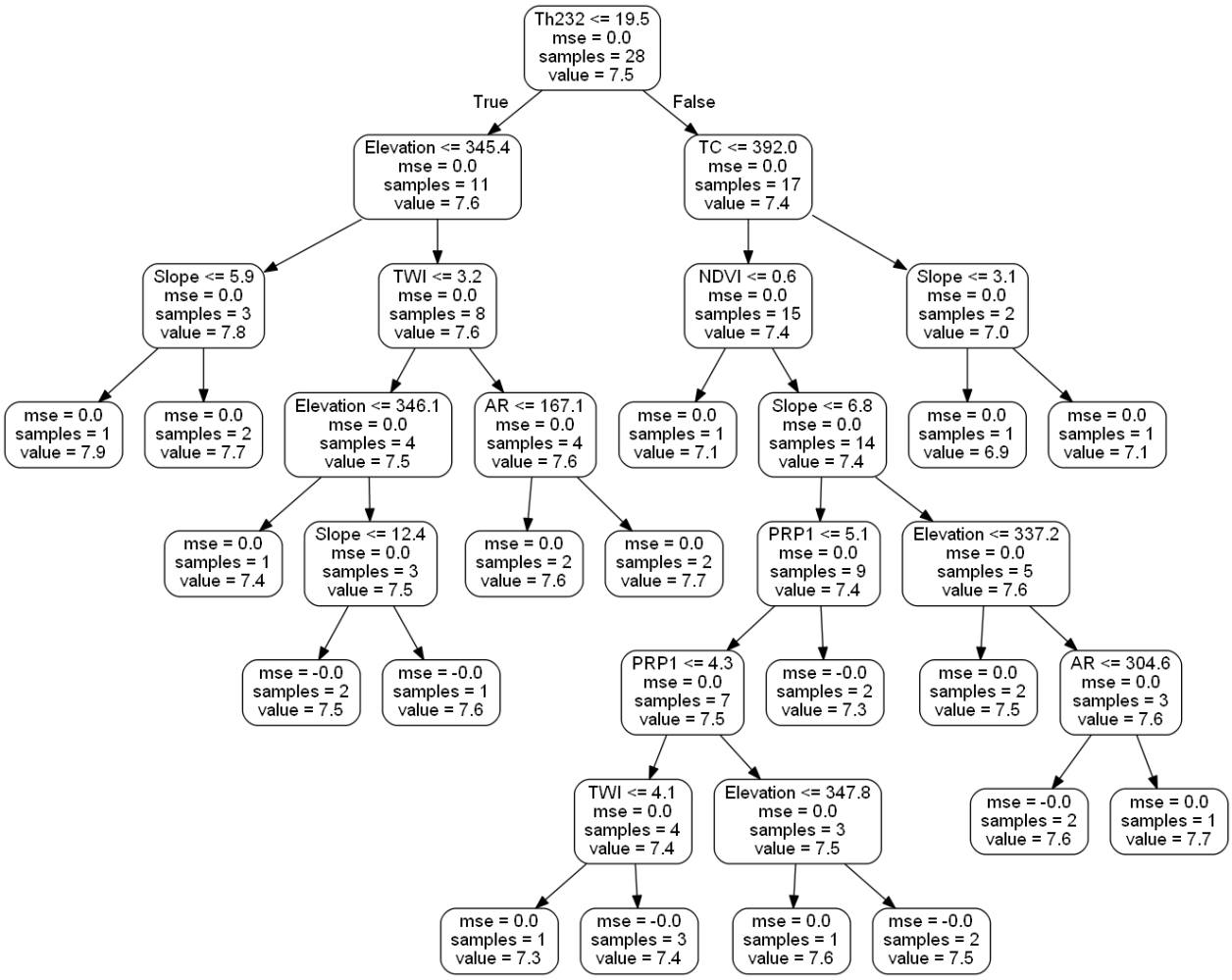



**Figure 5.6** Training (dataset split) and minimization of the node variance in the random forest model, an example for soil pH prediction.

In this study, the forest model used 70% of the entire datasets to train the model and 15% was used to undertake cross-validation (CV) or test, and the remaining 15% served for the final validation of the assigned predicted class. In the training phase, of the 70% of the entire sample datasets, the model used 65% (the unique dataset) to build the trees, while the remaining 35% were used for internal testing. The CV features of the RF algorithms improve the performance of the model while using independent test data (Blanco *et al.*, 2018). In the present study, five-fold cross-validation ($k = 5$) techniques for determining the final parameters of the model were tested after the training step. The tree model was tuned through CV procedures using a fine search grid. In considering the fit of a model with $k = 5$ in the test stage, the model used 80% of the independent dataset (from 15% of the test data) in each fold, and the remaining 20% of the data served in

estimating the predictive accuracy using the regression function. This step generated multiple train-test splits to tune the model. Finally, the accuracy was estimated based on the average performance on each fold (Zhou *et al.*, 2019).

In the training phase, the model was assessed with a different combination of samples (*random_state*) to create the regression tree. At the initial stage of the random forest model, the user needed to define *random-state* values where it selects the same combinations in each run and produces the same training/test data points to be run multiple times. Otherwise, the model produces different results (if it is fully random) in every run.

#### 5.2.8.3 Variable importance and optimization

A key step in building the prediction model, variable importance and ranking of the dataset were rigorously assessed. The RF model orders influential variables based on either mean decrease in prediction accuracy, or homogeneity (assess the quality of each variable split) of a variable split in the successive nodes (Heung *et al.*, 2014). The subset of the variable is determined by how the tree-based regression fits the soil prediction (Hengl *et al.*, 2018). In this study, the variable importance plot was derived from the RF default settings (*max_features* = none, where all features or variables are considered in each split instead of a random subset). At the training phase, a predicted value is assigned by determining the mean error rate and by averaging the predictions from the individual regression trees. After the removal of less important predictor variables, a revised soil prediction was performed to compare with the predicted results using all the variables and the reduced number of variables. In this study, a regression function was used for assessing the performance at the cross-validation stage.

#### 5.2.8.4 Model evaluation

The algorithms provide some solutions to enhance random forest optimization. A validation subset of grid samples was assessed based on the prediction results for measured soil values. The prediction accuracy depends on many accurate datasets (measured values) and many training samples. One of the most important advantages of many variables is that it reduces unintended model overfit. The root mean squared error (RMSE) was calculated for assessing model uncertainty, and coefficient of determination ($R^2$) assessed the degree of relationship

between predicted and measured values. The RMSE was used for the performance of the soil prediction results and it is described in the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}\left(y_{p_i} - y_{m_i}\right)^2} \quad (7)$$

where, *n is* the number of observations, $y_{m_i}$ is the $i^{th}$ measured value, and $y_{p_i}$ is the $i^{th}$ predicted value. The RMSE measures expected deviation of predicted values from their measured values.

## 5.3 Results

### *5.3.1 Descriptive analysis of the soil measurements*

Ranges between maximum and minimum values for the soil properties were variable throughout the whole field data and validation dataset (Table 5.6). In this study, the range (minimum and maximum values) of each soil property in the validation dataset was not always identical to the range of the whole dataset. The range, standard deviation (σ), and mean (μ) for each soil parameter showed large variability in the whole field (as described in Section 5.2.4) and validation dataset. In the validation dataset, the measured pH values varied between 6.8 and 7.7 (σ = 0.28 and μ = 7.4). For the SOM measurements, the range varied between 1.9% and 4.6% in the whole dataset (σ = 0.51%), whereas it varied between 2.9% and 4.2% in the validation dataset (σ = 0.46%). In the validation set, CEC measurements varied between 11 and 14 meq $hg^{-1}$ (between 9 and 18 meq $hg^{-1}$ in the whole set). The standard deviation value of Ca measurements varied largely between the two datasets (σ = 276.01 in the whole set; σ = 191.60 in the validation set). Moderate variability was found in P measurements (μ = 36.90 ppm in the whole set; μ = 34 ppm in the validation set), while K showed high variability (μ = 183.79 ppm in the whole field; μ = 205.30 ppm in the validation set.

**Table 5.6** The descriptive statistics of soil property values obtained through whole and validation sample dataset.

| Soil properties | Whole dataset | | | | Validation dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | **Min** | **Max** | **Mean** | **STD** | **Min** | **Max** | **Mean** | **STD** |
| pH | 6.50 | 7.90 | 7.44 | 0.26 | 6.80 | 7.70 | 7.40 | 0.28 |
| SOM (%) | 1.90 | 4.60 | 3.38 | 0.51 | 2.90 | 4.20 | 3.33 | 0.46 |
| P (ppm) | 15.00 | 70.00 | 36.90 | 12.24 | 22.00 | 52.00 | 34.00 | 8.91 |
| K (ppm) | 79.00 | 352.00 | 183.79 | 55.87 | 139.00 | 262.00 | 205.30 | 48.25 |
| Mg (ppm) | 123.00 | 393.00 | 267.66 | 63.77 | 160.00 | 370.00 | 267.20 | 72.40 |
| Ca (ppm) | 1236.00 | 2995.00 | 1741.89 | 276.01 | 1412.00 | 2086.00 | 1735.20 | 191.60 |
| Zn (ppm) | 1.50 | 17.00 | 5.98 | 3.69 | 1.80 | 10.70 | 4.95 | 2.84 |
| Mn (ppm) | 12.60 | 25.00 | 17.19 | 2.73 | 12.60 | 20.10 | 16.92 | 2.59 |
| CEC (meq $hg^{-1}$) | 9.00 | 18.00 | 12.58 | 1.52 | 11.00 | 14.00 | 12.60 | 0.97 |

*5.3.2 Analysis of correlation between high-density data and soil properties measured in the lab*

According to the relationship between the predictor variables (sensor variables and sensor-derived variables) found in Figure 5.4, most of the variables were considered for the prediction model. Pairwise relationships between the sensor variables and the soil properties and their strengths are shown the correlation matrix in Figure 5.7.

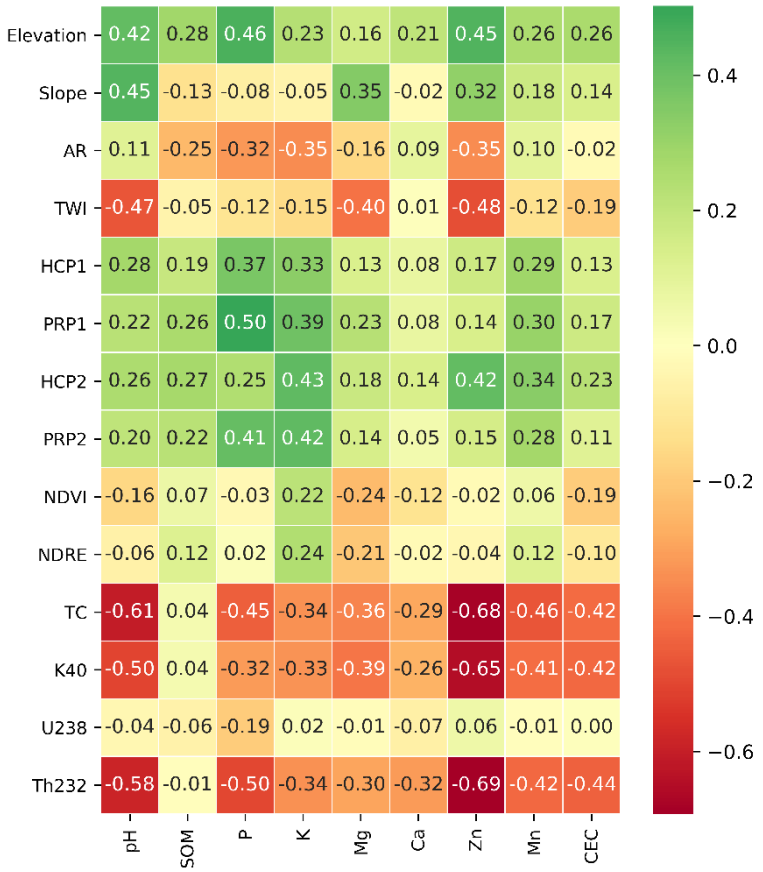



**Figure 5.7** Correlogram showing the relationship between predictor variables and different soil properties. The intensity of the green to red color increases with higher positive and negative correlation values.

For the experimental field, the pH correlated positively ($r$ = 0.42 and $r$ = 0.45) with the topographic variables of elevation and slope, respectively. Soil phosphorus (P) also showed a positive correlation with $EC_a^{0.0.5}$ and was negatively correlated with $^{232}$Th (r = -0.50). K was moderately well correlated with $EC_a$ variables ($0.33 \leq r \leq 0.43$) and with gamma sensor variables (r = -0.50) and, whereas Mg was moderately correlated with TWI (r = -0.40) and K40 gamma nuclide (-0.39). However, there was a negative correlation with the gamma-ray sensor, TC ($r$ = -0.61) and $^{232}$Th ($r$ = -0.58) for predicting pH. Cation exchange capacity (CEC) and Manganese (Mn) both correlated negatively with all gamma-ray sensor variables (except $^{238}$U). All gamma-ray variables showed a strong negative correlation for Zn prediction. No systematic correlations for SOM, K, Mg and Ca were found with any sensor measurements. Given the multifaceted linear relationship between a large number of predictor variables and the targeted

variables, it was challenging to evaluate the prediction capabilities of the sensor measurements for a specific soil property.

*5.3.3 Parameter optimization and variable reduction in RF*

In the RF regression model, from the original data (62 groups of data) of soil properties, 70% of the dataset was randomly divided into training data and 15% into both to serve as a test set and a final validation set. About 70% of the training data were randomly selected (see Section 5.2.8) for developing the forest model estimators and evaluating the parameters in the training model. About 15% of the data were selected for cross-validation and performance evaluation of the regression model estimator at this initial stage. In this study, approximately one-sixth of the data points (10 out of 62 sample datasets) served for the final validation and accuracy assessment of the regression models.

The targeted soil properties were pH, SOM, P, K, CEC, Mg, Mn, Zn, and Ca (see Section 5.2.4). Input parameters in the regression tree model are the predictor variables ($n$ = 14) that have a different effect in each soil prediction result. Details of the construction procedures of the predictive model and its application to the test data (unknown dataset) are shown in Figure 5.5. After the training and test data separation, optimization of different hyperparameters (mainly the number of trees and number of input variables) was required for the construction of each soil property model. At the training stage, the training dataset evaluated different combinations of sensor variables to fit a regression model and determine the parameters of the random forest model. The OOB error rates calculated from the RF model internal validation (outlined in Section 5.2.8). In the cross-validation stage (five-fold CV procedure), the $R^2$ determined optimum number of trees (*n_estimators*) for the model (Figure 5.8). The *n-estimators* values were selected from a range of 50-1000 in the trained model. pH curve was flat after 250, whereas the SOM, CEC, and Ca curve increased dramatically when the *n-estimators* was 50 (reached the maximum height at 150) and then leveled off where the *n-estimators* value was 300. For P and K prediction, the $R^2$ value was initially improved at 50 but decreased with the increase of *n-estimators* values. Mn and Zn prediction curves were increased until 250 and then leveled off. The optimum value for *n-estimators* with a five-fold CV procedure was within a range from 50 to 250 for the different soil

properties. Based on the initial results, the optimum value of *n-estimators* was 100 where $R^2$ was at the maximum for all soil predictions.

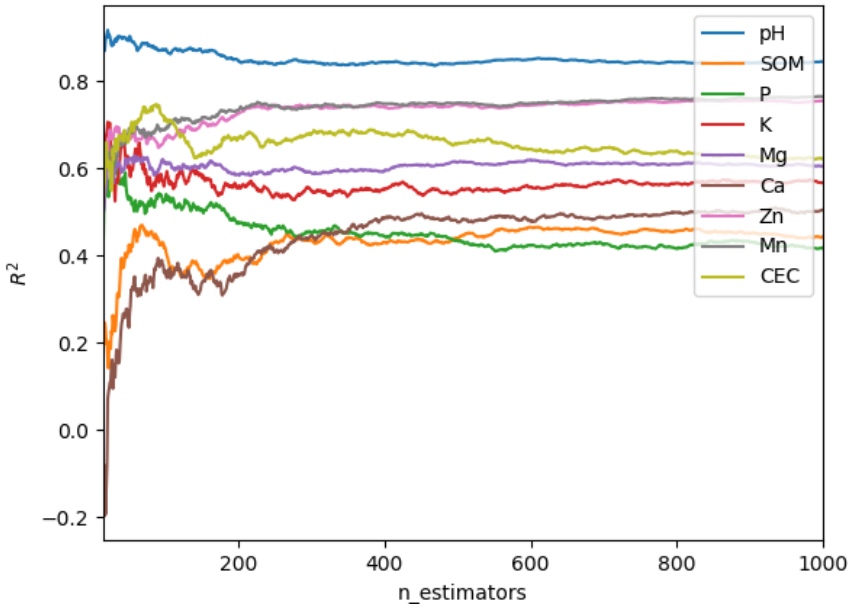
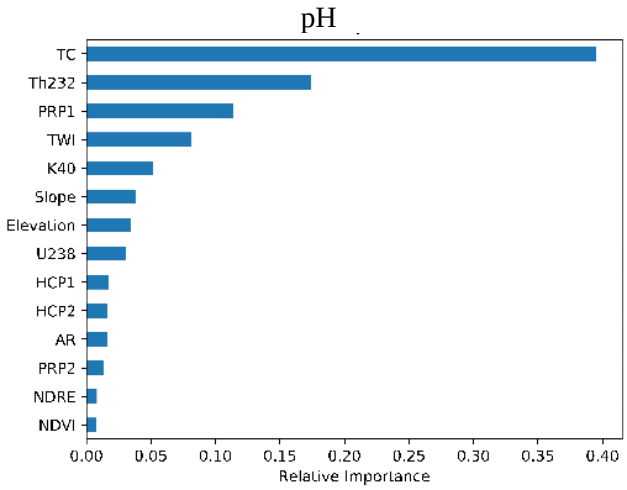
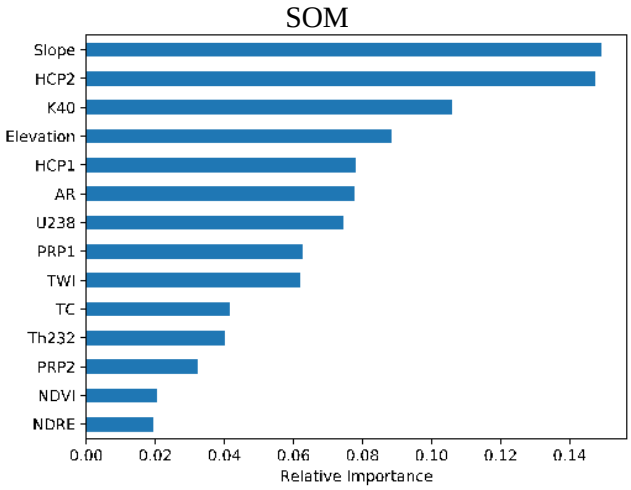
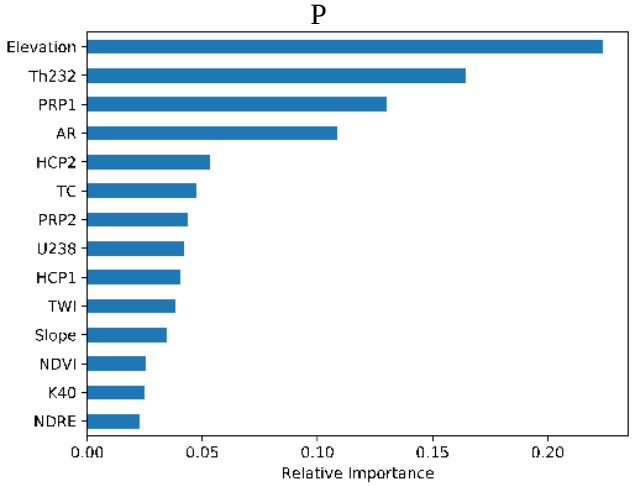


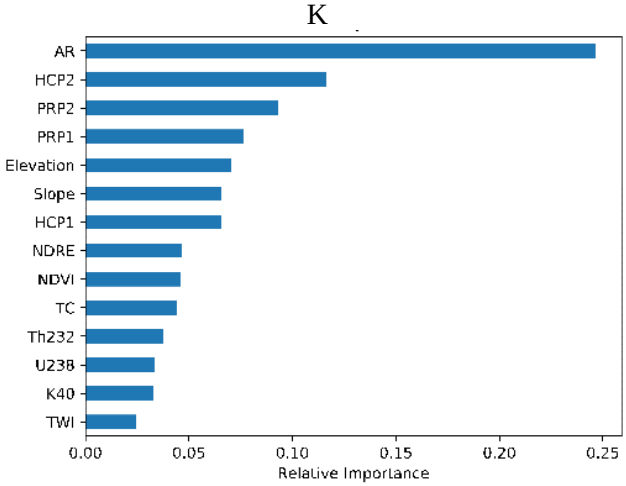

**Figure 5.8** Number of trees (*n_estimators*) optimizing for nine soil properties prediction. The coefficient of determination ($R^2$) has increasing trends at the cross-validation stage when *n_estimators* value between 50 and 100.

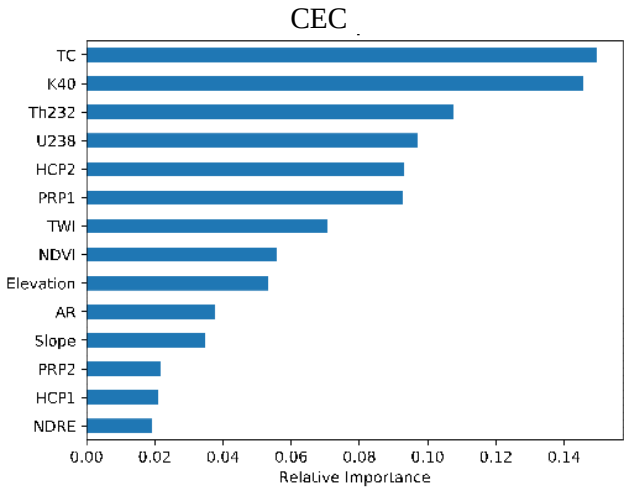
The sensitivity analysis of individual variables was evaluated by the degree of contribution when the RF model split a node in making the decision. In this study, a single approach variable reduction (default settings) was tested. Figure 5.9 shows the relative variable importance, when all four sensors ($EC_a$, topographic and gamma-ray, and satellite image) variables were considered for predicting all soil properties. The RF model evaluated the relative importance of 14 variables. Less influential variables were removed manually for testing the model's performance. After several runs, $R^2$ reached the maximum level in the independent cross-validation phase when the number of dominant variables were selected by the user based on their relative importance. In Table 5.7, the number of influential variables varied (ranging from high to moderate, from 3 to 11) for the prediction model of each soil property until the maximum $R^2$ value was achieved. This result was comparable with the results of obtaining a higher correlation coefficient (as discussed in Section 5.3.2). In this research, there does not appear to be a magic number of variables for all prediction models. The performance of the different combinations of variables affected the overall performance of the model. The overall performance of the selection is reported in a later section (through final validation).

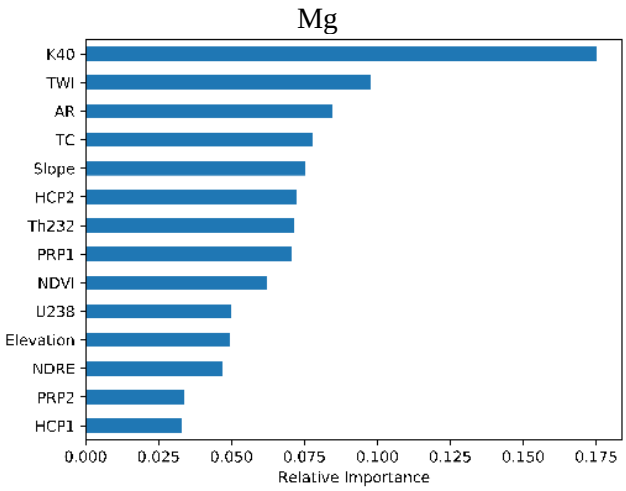

pH
TC
Th232
PRP1
TWI
K40
Slope
Elevation
U238
HCP1
HCP2
AR
PRP2
NDRE
NDVI
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Relative Importance


SOM
Slope
HCP2
K40
Elevation
HCP1
AR
U238
PRP1
TWI
TC
Th232
PRP2
NDVI
NDRE
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
Relative Importance


P
Elevation
Th232
PRP1
AR
HCP2
TC
PRP2
U238
HCP1
TWI
Slope
NDVI
K40
NDRE
0.00 0.05 0.10 0.15 0.20
Relative Importance


K
AR
HCP2
PRP2
PRP1
Elevation
Slope
HCP1
NDRE
NDVI
TC
Th232
U238
K40
TWI
0.00 0.05 0.10 0.15 0.20 0.25
Relative Importance


CEC
TC
K40
Th232
U238
HCP2
PRP1
TWI
NDVI
Elevation
AR
Slope
PRP2
HCP1
NDRE
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
Relative Importance


Mg
K40
TWI
AR
TC
Slope
HCP2
Th232
PRP1
NDVI
U238
Elevation
NDRE
PRP2
HCP1
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175
Relative Importance

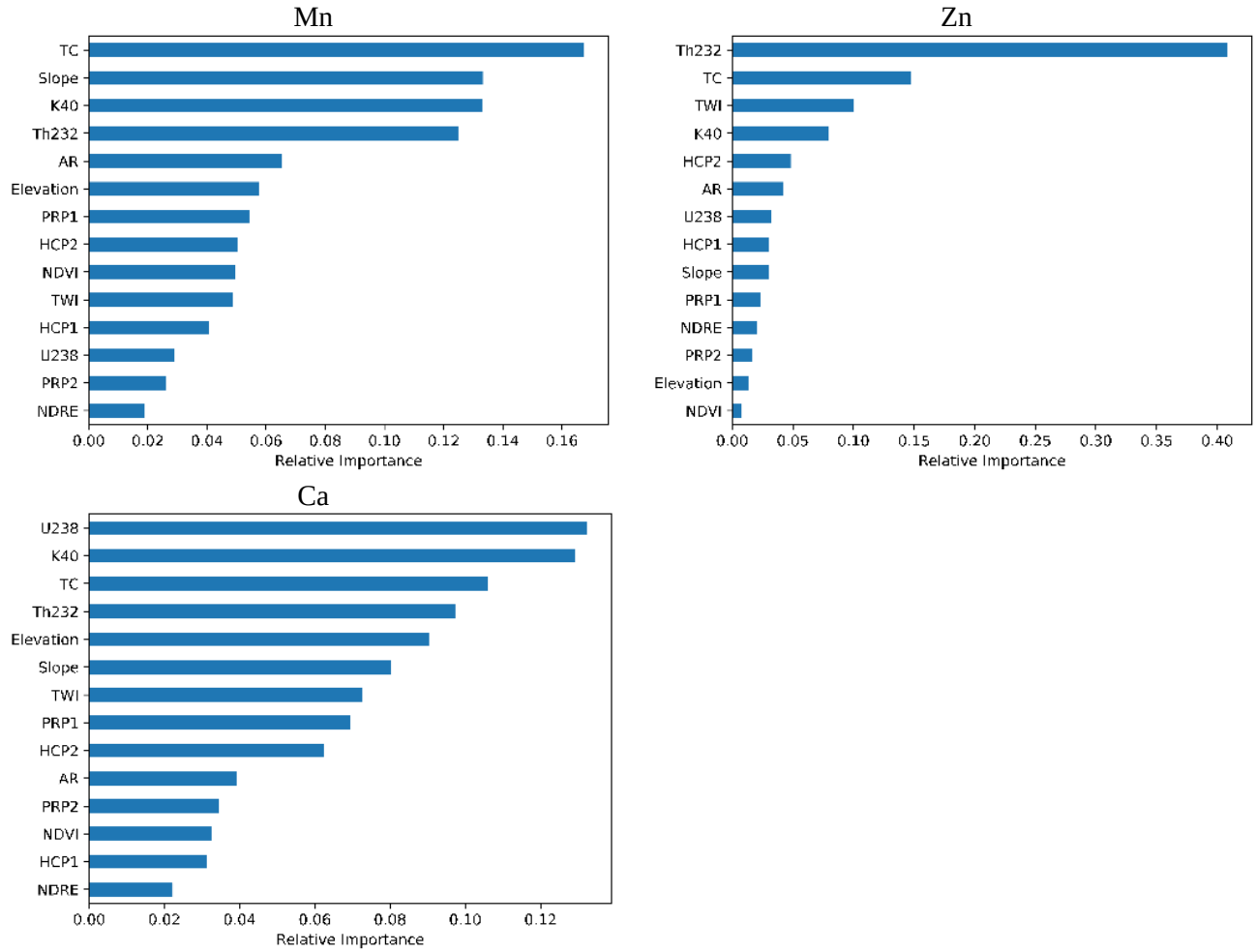



**Figure 5.9** Relative importance of the variables (derived from combining the four sensor's variables) for predicting nine soil properties in the random forest model.

Among the four $EC_a$ variables, shallow $EC_a$ (PRP1: 0-0.5 m) along with deep $EC_a$ (HCP2: 0-3.2 m) were most influential in predicting all soil properties. Shallow $EC_a$ were primarily affected many soil properties which are available for agricultural crop (Sudduth *et al.*, 2013). Among the topographic variables, elevation along with aspect ratio (AR) had a significant impact on constructing the RF prediction model. Two variables among γ-ray nuclides (TC and $^{232}$Th) were found to be important for building the soil prediction model. Between the two surface vegetation indices (derived from satellite image), NDVI was relatively more important than the NDRE for all of the prediction models. In most cases (Table 5.7), the four most dominant variables from the four different sensors were elevation, shallow $EC_a$, $^{40}$K, and NDVI for predicting soil properties at the local scale.

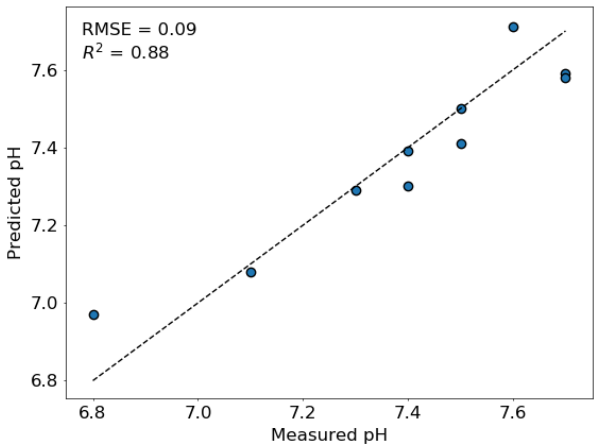
**Table 5.7** Optimum number of variables used in the final model based on the variable importance.

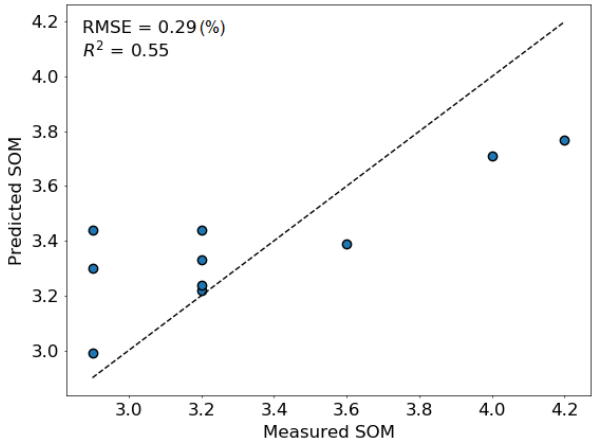
| Sensor variables | pH | SOM | P | K | CEC | Mg | Mn | Zn | Ca |
|---|---|---|---|---|---|---|---|---|---|
| Elevation | M | H | H | M | M | L | M | L | M |
| Slope | M | H | M | M | L | M | H | M | M |
| AR | L | M | H | H | L | H | M | M | L |
| TWI | H | M | M | L | M | H | M | H | M |
| $EC_a^{0\cdot1.6}$ | L | M | M | M | L | L | M | M | L |
| $EC_a^{0\cdot0.5}$ | H | M | H | M | M | M | M | M | M |
| $EC_a^{0\cdot3.2}$ | L | H | M | H | M | M | L | M | M |
| $EC_a^{0\cdot1.0}$ | L | L | M | H | L | L | M | L | L |
| TC | H | L | M | L | H | M | H | H | H |
| $^{40}$K | M | H | L | L | H | H | H | H | H |
| $^{232}$Th | H | L | H | L | H | M | H | H | H |
| $^{238}$U | M | M | M | L | H | L | L | M | H |
| NDVI | L | L | L | L | M | M | M | L | L |
| NDRE | L | L | L | L | L | L | L | M | L |

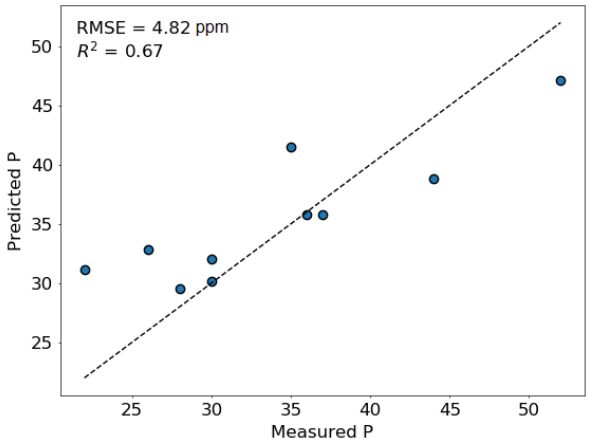
Note: H – high importance, M – moderate importance, and L – low importance

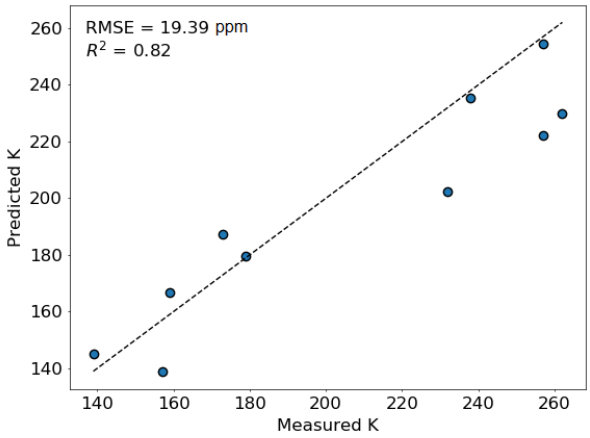
*5.3.4 Assessment of the prediction capability of the selected models*

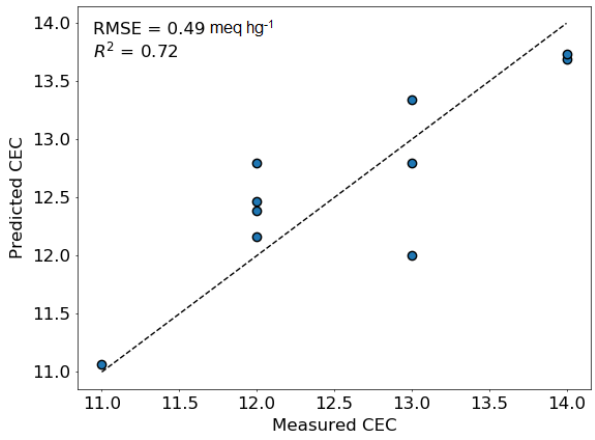
The performance of the different combinations of variables was assessed using the error of the prediction in the final validation step. Accuracy was assessed through the $R^2$ and RMSE (Figure 5.10). The actual *vs*. scatter plot showed that most of the soil prediction results were in close to perfect agreement (near 1:1 line), except SOM and CEC. A higher coefficient of the determination ($R^2 \geq 0.80$) was achieved in pH, K, and Zn predictions with the selected sensor variables (number of variables used: 8, 5 and 7, respectively). In this case, the estimated RMSE values were 0.09 for pH, 19.39 ppm for K, and 1.24 ppm for Zn. Sensor fusion required for CEC prediction included combining RTK with DUALEM and gamma-ray sensors, while Mg prediction combined gamma-ray and RTK sensors ($R^2 = 0.71$). Also, P prediction results were improved by combining only four variables ($R^2 = 0.67$). The SOM, Mn, Ca predictions were weaker ($0.50 < R^2 < 0.60$) when combined with multiple sensors — gamma-ray, RTK GNSS, and remote sensing sensors — which produced maximum prediction results from all other combinations.

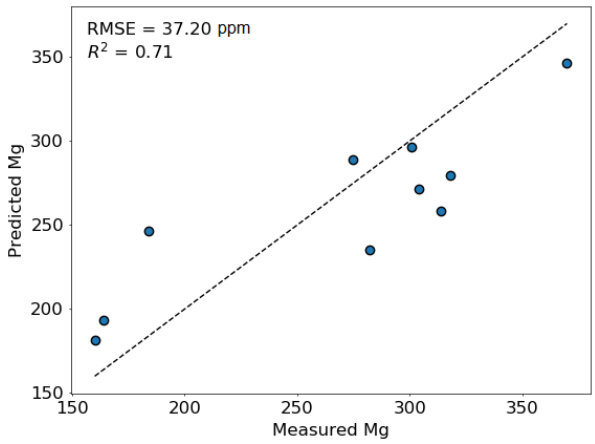

RMSE = 0.09
$R^2$ = 0.88
Predicted pH
Measured pH
6.8
7.0
7.2
7.4
7.6


RMSE = 0.29 (%)
$R^2$ = 0.55
Predicted SOM
Measured SOM
3.0
3.2
3.4
3.6
3.8
4.0
4.2


RMSE = 4.82 ppm
$R^2$ = 0.67
Predicted P
Measured P
25
30
35
40
45
50


RMSE = 19.39 ppm
$R^2$ = 0.82
Predicted K
Measured K
140
160
180
200
220
240
260


RMSE = 0.49 meq hg$^{-1}$
$R^2$ = 0.72
Predicted CEC
Measured CEC
11.0
11.5
12.0
12.5
13.0
13.5
14.0


RMSE = 37.20 ppm
$R^2$ = 0.71
Predicted Mg
Measured Mg
150
200
250
300
350

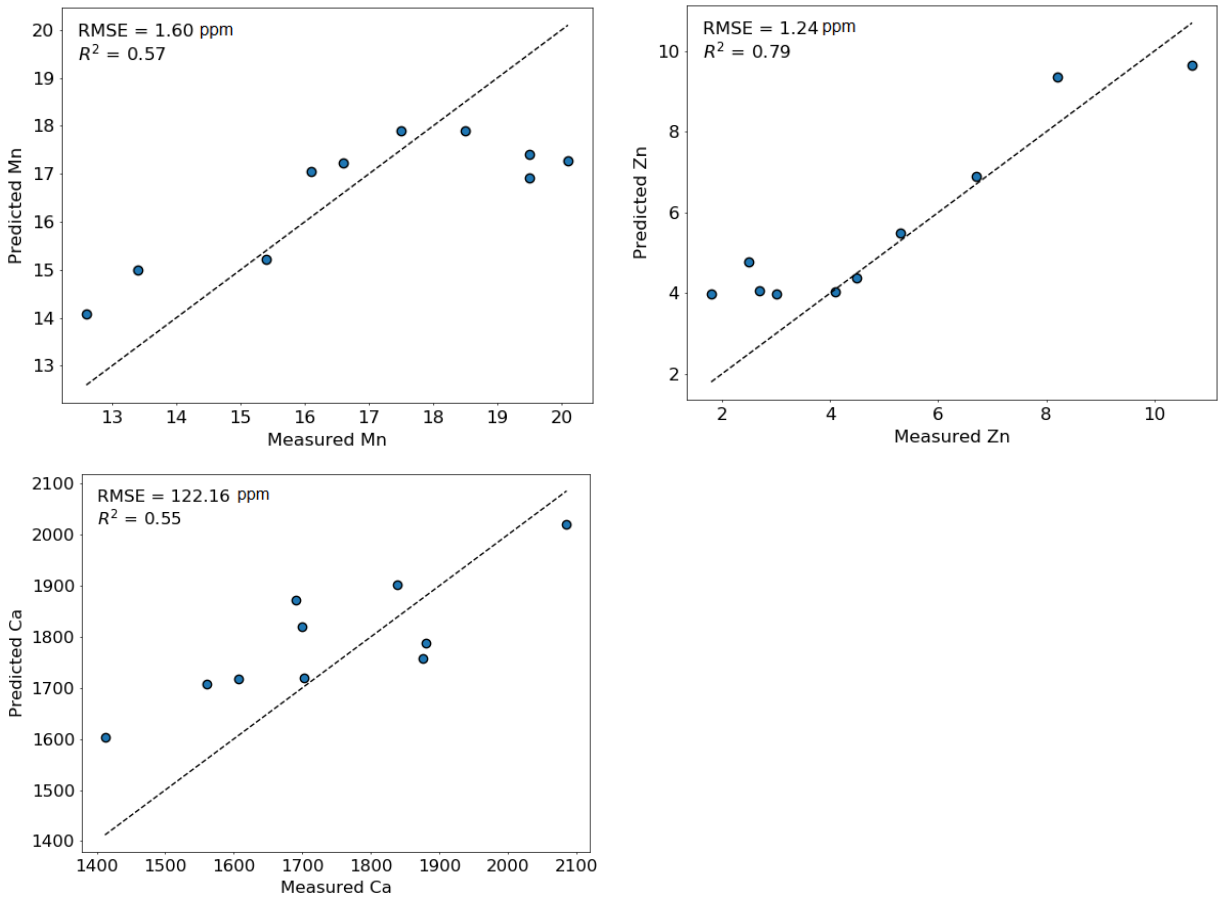



**Figure 5.10** Assessment of accuracy for prediction of various soil parameters – pH, Soil organic matter (SOM), Phosphorus (P), Potassium (K), Cation exchange Capacity (CEC), Magnesium (Mg), Manganese (Mn), Zinc (Zn), Calcium (Ca). Model accuracy evaluated using root mean squared error (RMSE) of each soil measurement, and coefficient of determination ($R^2$).

Most of the soil prediction results were within the accepted range of the measured soil samples (Figure 5.10). In all cases, the standard deviation of the predicted values was smaller than the measured values. The predicted mean values were lower than the mean value of measured soil properties (*i.e.*, pH, SOM, P, CEC, Mg, Mn, Zn).

## 5.4 Discussion

This study evaluated the complex relationship between sensor variables and soil physiochemical properties. Analysis of the data shows the negative and positive correlation for each soil property with auxiliary variables. Soil sensor data was collected from various platforms

at different times and from different soil depths. From the training datasets and the regression capabilities, the random forest model effectively predicted the above-mentioned nine soil properties. Tuning hyperparameters was the key to maximizing accuracy. By tuning a single hyperparameter (*i.e.*, number of regression trees), the model assessed a large number of auxiliary variables and their combinations for predicting all soil properties.

In this regression model, more data points are required to improve the model's performance. With only a few training datasets, they cannot represent all soil measurements from a large field and underfitting often occurs as a result of unknown data. On the other hand, too large a training dataset reduces the performance of the generalization in the tree model. Heung *et al.* (2016) determined that the optimum number of training and test (cross-validation) data sets was key to improving model precision on a regional scale. In the present study, larger numbers of trees perform better in training the model. However, building a lot of trees, splitting results in the training phase and then averaging the results of the regression trees can slow down the training process considerably. Therefore, the parameter search should find a sweet spot (optimal number) to increase the efficiency of the prediction model.

The topographic sensor (RTK GNSS) along with the $EC_a$ (DUALEM-21S) sensor and most outputs of the gamma-ray (SoilOptix®) sensor can be combined to predict soil nutrients at the field scale. It was found that the variable of low importance did not contribute much to the model. Sometimes, the coefficient of determination was improved significantly at the cross-validation stage when the low importance variables were removed from the model. For instance, the $R^2$ of Mn prediction was improved to 0.57 when only six variables were considered; however, it was decreased to 0.45 when all sensor variables were considered. Higher performance was observed in pH, K, Zn prediction using the least number of variables and with small datasets at the local scale. On the other hand, SOM, P, CEC, and Ca were estimated using most of the sensor variables to improve accuracy. Combining RTK with the gamma-ray sensor provided the best prediction results for pH, Mg, Mn, Zn, while combining gamma-ray and topographic sensors with the $EC_a$ sensor (DUALEM-21S) improved the model for SOM and CEC estimation. Remote sensing vegetation indices combined only with DUALEM were effective in predicting K, while indices combined with RTK sensor output were effective for Ca. Predicting P properly was more challenging and required a combination with all other sensors (including the gamma-ray sensor)

to predict P properly. Gamma-ray and topographic sensors combined estimated with a lower estimation error and achieved higher prediction results for all micro- and macro-nutrients in the agricultural field. This was explained by the fact that most of the subsurface soil chemical properties change with changes in the topographic parameters. The gamma radionuclides were efficient in detecting parent materials and soil texture, which generally determine soil physical properties (e.g., SOM, CEC, etc.). This indicates that soil variability determined using sensor-fused data and regression techniques could assist in constructing precise prediction models for soil properties and in developing site-specific crop management. One of the most important advantages of having a large number of variables is that it reduces the unintended model overfit.

This research relied on relatively small datasets (only 62 data points with a different combination of variables). The optimum precision of the model and generalization of estimators depends on the quality of the sensor measurements and on having many data points. In this study, a default feature selection or all variables considered each soil property prediction; however, it needs to address alternative approaches to determine the optimum number of variables for future research. Cross-validation techniques inside the forest model provided accuracy assessment of the trained model and in some cases, showed the model could not accurately fit the unknown data (Zhou *et al.*, 2019). Compared to the cross-validation steps for model performance, variable reduction is still less effective in improving the model accuracy. Due to the comprehensive sensitivity issues of the variables, the data, even though it was collected from the same area, may provide different results in other machine learning models. Texture data and the other physical properties would make the prediction method more robust. Also, the complex nature of the relationship between predictor variables and soil properties is often difficult to explain.

In many regions, micro- and macro-nutrient prediction is essential to understand soil variability in a large agricultural field (Mahmood *et al.*, 2012). High-density PSS data collected temporally along with lab-based measurements makes the model efficient in analyzing other environmental variables and in their prediction. The laboratory analysis data and precise thematic maps provide a better indication of field management and fertilizer recommendations. In some parts of the field which have lower pH, lime requirements for certain crops can be subjective. This is especially true if the field contains a lower amount of soil organic matter, which can be increased

by cover crops and mulching. This research may lead to the development of better thematic soil maps which can improve site-specific farm management techniques in the future.

## 5.5 Conclusions

This research investigated the prediction capability of different sensor variables using a random forest regression method for predicting nine soil properties in an agricultural field. Better results in predicting farm-scale soil properties were obtained through the integration of proximal and remote sensing sensor variables. The gamma-ray and RTK GNSS sensors were found to be the most valuable for soil parameter prediction and mapping at the local level. Freely available multispectral remote sensing data combined with gamma sensor variables can predict important soil properties. The regression tree model could assist in establishing a hierarchical relation between sensor variables, as well as efficiently selecting important variables with less user influence. The model demonstrated efficiency in terms of combining different sensor variables and identifying optimal values of input parameters. Error reporting at the earlier stage of the training phase and fewer user inputs make the supervised model robust in digital soil mapping. The model accuracy depends on the number of training samples and the optimum number of important variables selected. One of the most important advantages is that the use of many variables reduces the unintended model overfit. Internal model validation and cross-validation could increase accuracy and efficiency for the digital soil mapping process in other areas. Although this research used well-distributed 62 samples along with an independent validation dataset, experiments in different agricultural fields with more measurements would increase the acceptability of the model in other agricultural studies. Although direct and intensive soil measurements are a reliable method, they are an expensive and time-consuming procedure for crop production. This effort seeks to reduce the number of sample measurements while considering the important number of the sensor variables and increasing the understanding of field variability at the local scale.

The developed algorithm and model will improve soil prediction methods and provide tools for a decision support system in any dynamic production system. This research offers strategic opportunities and advantages for crop advisors to make faster decisions based on accurate soil mapping. Accurate mapping can also optimize the production system's profitability by reducing agricultural inputs and maximizing environmental benefits with the goal of sustainability.

# Chapter 6: Summary and General Conclusions

## 6.1 Summary

Having a large quantity of geospatial data having been collected using multiple proximal soil sensing (PSS) and remote sensing (RS) sensors facilitate soil characterization procedures for monitoring soil and crop management. By identifying the variability of different parts of a field, the current research first optimized field-based zonal homogeneity for model-based soil characterization. Then, high-density soil measurements were deployed to investigate the model's predictive ability for multiple soil properties. Finally, a machine learning method was developed to optimize the parameters of the geospatial data integration and estimate the prediction accuracy of the thematic mapping process, in an effort to create a precise digital soil map.

The first part of the present study examined the use of a hierarchical data clustering technique drawing on PSS and RS sensor-based soil responses to determine relatively homogeneous parts of agricultural fields. Multivariate data — (i) shallow and deep apparent soil electrical conductivity ($EC_a$), (ii) high-accuracy topographic indices, and (iii) bare soil and vegetation indices (VIs) — were collected from three agricultural fields in Ontario, Canada. The Neighborhood Search Analyst (NSA) data clustering tool's ability to define spatial continuity in zone delineation was assessed and used to characterize soil variability. The performance of this technique was found to be better than that of fuzzy clustering methods in producing the optimum (or user-defined) number of zones. These homogeneity maps provided field variability and essential information for monitoring and managing soil health in production fields. The field variability identified in this model arose from the successful optimization of PSS sensor-based zonal heterogeneity to achieve further agronomic model calibration. The information developed in this study will be essential to guide crop advisors who seek to optimize soil sampling locations and employ soil variability for the prediction and mapping of different variable rate applications.

After a rigorous assessment of the multiple variables and their zonal variability derived from the sensor response, a second study evaluated DUALEM-21S and RTK GNSS sensor-derived measurements against samples collected from targeted sample locations in a large number of agricultural farms operating under different agro-climatic conditions across Ontario, Canada. A

large quantity of high-density data ($EC_a$, and topographic indices) were obtained and multiscale field variability was analyzed within a statistical framework to optimize the soil prediction capability. This research explored sampling strategies by: (i) evaluating soil sensing measurements collected by two sensors and the quality of their data, and (ii) optimizing model prediction capacity for six selected soil properties. The measurement errors and prediction efficiency were assessed and compared to the median absolute deviation (MAD) values measured through the North American Proficiency Testing (NAPT) program. After assessing the sensor-based predicting efficiencies of the lab results, NAPT thresholds were used as a benchmark for evaluating accuracy. This could be potentially useful for standard laboratory citification programs. This study showed the powerful potential of proximal soil sensing technologies to predict soil nutrients and to allow mapping for site-specific crops and soil management in precision farming. This protocol of sensor data optimization can be used by commercial sensor users and agronomic service providers to improve their data handling processes and maximize the information value of the data they generate for their customers.

As part of the process of predicting targeted soil nutrients, the final portion of the project used a decision tree-based method to assess the model's prediction capacities and determine soil variability. A wide range of environmental covariates — (*i*) vegetation indices from multispectral remote sensing spectra, (*ii*) topographic indices from RTK GNSS, (*iii*) apparent soil electrical conductivity ($EC_a$) from DUALEM, and (*iv*) radionuclide variables from gamma sensors — were mapped in an agricultural field located in Ontario, Canada. A subset of sensor measurements and georeferenced soil sample data were used to predict multiple soil nutrients in the production field. Random forest algorithms were constructed to optimize model parameters at the training stage. A sensitivity analysis was performed to obtain the best results and scenario maps. The model has a unique capacity to optimize parameters while handling overfitting. Model performance was assessed by evaluating the prediction results of multiple soil nutrients with independent validation datasets. Soil variability determined using sensor-fused data and related techniques could assist in constructing precise prediction models for soil properties and in developing reliable thematic maps for field-scale crop management. Based on the arguments presented in the above discussion of sensor data optimization and modeling results, this research may lead to the development of better thematic soil maps and site-specific farm management techniques in the future.

## 6.2 General conclusions

The assessment of high-density multivariate data and soil characterization is one of several requirements to generate an accurate soil map for use in precision farming. Our multivariate geospatial data mining models play a key role in soil prediction and digital mapping processes. A hierarchical data analysis model provided a unique field variability map and stabilizing information for optimizing soil sample measurements. The preprocessing and variable selection steps common to all clustering techniques are imperative for providing a delineated areal extent (DAE) for developing thematic maps. Compared to other data clustering algorithms, the NSA clustering tool showed a unique capacity to provide spatially-contiguous clusters, allowing the delineation of an optimum number of zones. Moreover, this software was tested and demonstrated that it was capable of handling a significant number of variables and high-density data layers for delineating the optimum (or defined) number of zones in a more precise way. The robust zone delineation process and georeferenced thematic maps increase efficiency for variable-rate crop management technologies and are useful for other management purposes.

This research optimized models by assessing proximal soil sensor data (high-density apparent electrical conductivity and topographic indices) and their predictive properties for the determination of soil nutrients. Topographic variables showed promising results for the prediction of soil organic matter and CEC in agricultural fields in Ontario, Canada. Shallow $EC_a$ plays an important role in understanding within-field variability; however, evidence of the applicability of tested proximal sensing technologies to address spatial variability of certain soil nutrients, such as potassium (K) proved to be rather limited. In another part of this study, the topographic indices, $EC_a$ parameters along with gamma radionuclides and vegetation responses were modeled to achieve the best prediction results for multiple soil properties in a production field. In the present study, a decision tree-based model was applied to determine the importance of each variable. In the prediction model, an optimum number of variables, mainly topographic, gamma nuclides and typically normalized difference vegetation index, were employed to achieve the best prediction results for several properties (*i.e.*, pH, K, Zn), while only the remote sensing vegetation index combined with $EC_a$ data were effective in predicting K at the field-scale. The random forest (RF) regression (training and testing) analysis indicated that soil variability determined using sensor-fused data and methods provided better assistance in the construction of precise prediction models

for macro- and micro-nutrients (*i.e*., pH, K, CEC and Zn). Performed for the predicted soil properties using small datasets to evaluate the model's predictive accuracy, the modeling processes were more effective in developing reliable digital soil maps than traditional statistical models.

Our findings indicate the powerful potential of proximal soil sensing technologies to define the site-specific crop production environment in terms of terrain and soil physical characteristics. The present results suggest that the integration of conceptually different sensors for multiple soil measurements is useful in optimizing soil characterization and allows a better prediction of certain soil properties than a single type of measurement. Furthermore, it also improves the soil thematic maps. Without using these precision technologies and methods, it is quite challenging to deal with multiscale optimization in an heterogenous landscape or production system, and almost impossible to produce a precise digital map. Optimized sampling and erroneous data removal models, supervised machine learning prediction frameworks for high-density geospatial data, could be implemented as web applications to facilitate appropriate site-specific agronomic and environmental decisions. Continuing research efforts will explore additional measurement capabilities that could potentially expand the applicability of proximal soil sensing tools.

Moreover, this research may lead to the development of better thematic soil maps that can improve digital soil mapping techniques and future site-specific farm management approaches, thereby, increasing the probability of making the landscape profitable and environmentally sustainable. These soil maps can be used to implement variable-rate fertilizer recommendations, liming, or seeding density, thereby, optimizing the use of agricultural inputs by crop producers, their consultants, and agribusiness representatives. A scaled-up adoption of proximal soil sensing technologies would provide advances in agricultural crop production, sustainable resource management and provide great environmental benefits in Canada and the rest of the world.

# Chapter 7: Contribution to Knowledge and Suggestions for Future Research

## 7.1 Contribution to knowledge

The current research generates knowledge on the processing of high-density and sensor-fused data at different geospatial scales as well as providing more information on soil thematic mapping. The newly developed hierarchical data analysis software handled multi-dimensional variables for understanding zonal heterogeneity and various management practices. This tool can significantly contribute to big data mining processes in precision farm management. Furthermore, multiscale soil characterization combining surface and subsurface information provides a unique guideline for crop producers. This study offers essential knowledge on retrieving and analyzing high-density sensor data to achieve cost-effective soil sampling and sensor-based soil nutrient estimation.

The tangible contribution is the ability to evaluate similar data for many agricultural fields across Ontario. Also, this research is unique in combining RS and PSS data for many fields. Evaluating $EC_a$ and gamma-ray data in parallel is new in terms of exploring soil variability at the farm scale. This research explored the elements of advanced data modeling, such as the regression forest. The results did not indicate strong predictability for some chemical properties, which contribute to understanding agronomic properties, and consequently, innovative analysis is needed for an improved understanding of soil heterogeneity to enhance the efficiency of site-specific crop management. This can be done by looking at the differentiation of seeding rates, irrigation/drainage and/or N management; however, this was outside the scope of this study.

The data integration algorithm and optimization of model hyperparameters improve the performance of soil prediction methods and provide tools for both local and regional scale decision support systems. The tree-based regression method and its thematic maps are very effective for farm scale soil variability assessment and faster decision making. In general, this research provides strategic opportunities to obtain precise thematic maps and to provide advantages for crop producers to enhance their decision making to ensure that the production system is profitable and the landscape remains sustainable over the long term. Ultimately, this will provide information for better variable-rate fertilizer recommendations and optimal pesticide and herbicide applications.

## 7.2 Suggestions for future research

Many critical questions in agricultural research are far from being solved. The present research with only three sets of objectives was able to address some of these issues; however, future work is recommended. Continuing research will need to explore additional measurements of soil physical properties with advanced soil sensing technologies. These tools provide an assessment of soil health and determine how it can be improved with amendments (manure, compost, cover crops, fertilizer, etc.). Moreover, future research will validate and implement results through a set of case studies and disseminate findings among the agricultural farming communities.

Moreover, sensor fusion with multi-temporal airborne (low and medium altitude platforms) image spectra, which is valued in many earth science applications, may offer an optimum solution for field-scale precision agriculture problems. Proximal soil sensing data combined with high-spatial and multi-temporal microwave data from the Canadian RADARSAT Constellation Mission (RCM) has not yet been explored for solving agricultural problems. As a big data source, the Google Earth Engine API will be a potential resource in integrating multi-temporal images with soil and environmental datasets for agricultural applications. Also, the integration of field-measured (or lab-measured) spectra with hyperspectral satellite spectra would be beneficial for digital soil mapping activities at the local level.

## References


1. Adamchuk, V.I., Lund, E.D., Sethuramasamyraja, B., Morgan, M.T., Dobermann, A., Marx, D.B. Direct measurement of soil chemical properties on-the-go using ion-selective electrodes. *Computers and Electronics in Agriculture*, **2005**, 48(3), 272-294.

2. Adamchuk, V.I.; Hummel, J.W.; Morgan, M.T.; Upadhyaya, S.K. On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture* **2004**, *44*, 71-91, doi:https://doi.org/10.1016/j.compag.2004.03.002.

3. Adamchuk, V.I.; Tremblay, N. New developments in proximal soil sensing. In Proceedings of The International Tri-Conference for Precision Agriculture, Hamilton, October 16, 2017.

4. Adamchuk, V.I.; Viscarra Rossel, R.A. Development of On-the-Go Proximal Soil Sensor Systems. In *Proximal Soil Sensing*, Rossel, R.A.V., McBratney, A.B., Minasny, B., Eds. Springer Netherlands: Dordrecht, 2010; https://doi.org/10.1007/978-90-481-8859-8_2 pp. 15-28.

5. Adamchuk, V.I.; Viscarra Rossel, R.A.; Marx, D.B.; Samal, A.K. Using targeted sampling to process multivariate soil sensing data. *Geoderma* **2011**, *163*, 63-73, doi:https://doi.org/10.1016/j.geoderma.2011.04.004.

6. Albornoz, E.M.; Kemerer, A.C.; Galarza, R.; Mastaglia, N.; Melchiori, R.; Martínez, C.E. Development and evaluation of an automatic software for management zone delineation. *Precision Agriculture* **2018**, *19*, 463-476, doi:10.1007/s11119-017-9530-9.

7. Alchanatis, V.; Cohen, Y. Special issue on sensors in agriculture. *Biosystems Engineering* **2013**, *114*, doi:10.1016/j.biosystemseng.2013.01.007.

8. Aldabaa, A.A.A.; Weindorf, D.C.; Chakraborty, S.; Sharma, A.; Li, B. Combination of proximal and remote sensing methods for rapid soil salinity quantification. *Geoderma* **2015**, *239-240*, 34-46, doi:10.1016/j.geoderma.2014.09.011.

9. Ali, A.M.; Thind, H.S.; Varinderpal, S.; Bijay, S. A framework for refining nitrogen management in dry direct-seeded rice using GreenSeeker™ optical sensor. *Computers and Electronics in Agriculture* **2015**, *110*, 114-120, doi:https://doi.org/10.1016/j.compag.2014.10.021.

10. Arabie, P.; Hubert, L.J. An Overview of Combinatorial Data Analysis. In *Clustering and Classification*; Arabie, P., Soete, G.D., Hubert, L.J., Eds.; World Scientific Pub. Co.: Singapore, 1996; pp. 5–63.

11. Asher, J.B.; Yosef, B.B.; Volinsky, R. Ground-based remote sensing system for irrigation scheduling. *Biosystems Engineering* **2013**, *114*, 444-453, doi:https://doi.org/10.1016/j.biosystemseng.2012.09.002.

12. Baldoncini, M.; Albéri, M.; Bottardi, C.; Chiarelli, E.; Raptis, K.G.C.; Strati, V.; Mantovani, F. Biomass water content effect on soil moisture assessment via proximal gamma-ray spectroscopy. *Geoderma* **2019**, *335*, 69-77, doi:10.1016/j.geoderma.2018.08.012.

13. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* **2016**, *114*, 24-31, doi:https://doi.org/10.1016/j.isprsjprs.2016.01.011.

14. Berkhin, P. A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data*; Springer: Berlin/Heidelberg, Germany, 2006; pp 25–71.

15. Beven, K.J.; Kirkby, M.J. A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrol. Sci. Bull.* **1979**, *24*, 43–69.

16. Bishop, T.F.A.; Minasny, B. Digital soil-terrain modelling: the predictive potential and uncertainty. In *Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics*, Grunwald, S., Ed. CRC Press / Taylor & Francis Group: Baco Raton, FL, 2006; pp. 185-213.

17. Blanco, C.M.G.; Gomez, V.M.B.; Crespo, P.; Ließ, M. Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest. *Geoderma* **2018**, *316*, 100-114, doi:10.1016/j.geoderma.2017.12.002.

18. Borgogno-Mondino, E.; Lessio, A.; Tarricone, L.; Novello, V.; de Palma, L. A comparison between multispectral aerial and satellite imagery in precision viticulture. *Precision Agriculture* **2018**, *19*, 195-217, doi:10.1007/s11119-017-9510-0.

19. Bragato, G. Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain. *Geoderma* **2004**, *118*, 1-16.

20. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5-32, doi:10.1023/a:1010933404324.

21. Brogi, C.; Huisman, J.A.; Pätzold, S.; von Hebel, C.; Weihermüller, L.; Kaufmann, M.S.; van der Kruk, J.; Vereecken, H. Large-scale soil mapping using multi-configuration EMI and supervised image classification. *Geoderma* **2019**, *335*, 133-148, doi:10.1016/j.geoderma.2018.08.001.

22. Brown, J.D. A Historical Perspective on Soil-Landscape Modeling. In *Environmental Soil-Landscape Modeling Geographic Information Technologies and Pedometrics*, Grunwald, S., Ed. Taylor & Francis: New York, 2006; pp. 61-103.

23. Burrough, P.A.; van Gaans, P.F.M.; Hootsmans, R. Continuous classification in soil survey: spatial correlation, confusion and boundaries. *Geoderma* **1997**, *77*, 115-135, doi:https://doi.org/10.1016/S0016-7061(97)00018-9.

24. Castrignano, A.; Buttafuoco, G.; Quarto, R.; Vitti, C.; Langella, G.; Terribile, F.; Venezia, A. A Combined Approach of Sensor Data Fusion and Multivariate Geostatistics for Delineation of Homogeneous Zones in an Agricultural Field. *Sensors (Basel)* **2017**, *17*, doi:10.3390/s17122794.

25. Cherlinka, V. Using Geostatistics, DEM and Remote Sensing to Clarify Soil Cover Maps of Ukraine. In *Soil Science Working for a Living*, D., D., Y., D., Eds. Springer: Cham, 2017; https://doi.org/10.1007/978-3-319-45417-7_7.

26. Cohen, S.; Cohen, Y.; Alchanatis, V.; Levi, O. Combining spectral and spatial information from aerial hyperspectral images for delineating homogenous management zones. *Biosystems Engineering* **2013**, *114*, 435-443.

27. Córdoba, M.; Bruno, C.; Costa, J.; Balzarini, M. Subfield management class delineation using cluster analysis from spatial principal components of soil variables. *Computers and Electronics in Agriculture* **2013**, *97*, 6-14, doi:https://doi.org/10.1016/j.compag.2013.05.009.

28. Córdoba, M.A.; Bruno, C.I.; Costa, J.L.; Peralta, N.R.; Balzarini, M.G. Protocol for multivariate homogeneous zone delineation in precision agriculture. *Biosystems engineering* **2016**, *143*, 95-107, doi:10.1016/j.biosystemseng.2015.12.008.

29. Corwin, D.L.; Lesch, S.M. Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture* **2005**, *46*, 11-43, doi:10.1016/j.compag.2004.10.005.

30. Cressie, N.; Kang, E.L. High-Resolution Digital Soil Mapping: Kriging for Very Large Datasets. In *Proximal Soil Sensing*, Viscarra Rossel, R.A., McBratney, A.B., Minasny, B., Eds. Springer: Dordrecht, 2010; 10.1007/978-90-481-8859-8_4pp. 49-63.

31. Dao, T.H. Sensing soil and foliar phosphorus fluorescence in Zea mays in response to large phosphorus additions. *Precision Agriculture* **2017**, *18*, 685-700, doi:10.1007/s11119-016-9480-7.

32. De Benedetto, D.; Castrignano, A.; Diacono, M.; Rinaldi, M.; Ruggieri, S.; Tamborrino, R. Field Partition by Proximal and Remote Sensing Data Fusion. *Biosyst. Eng.* **2013**, *114*, 372–383.

33. De Gruijter, J.J.; Walvoort, D.J.J.; van Gams, P.F.M. Continuous soil maps - a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma* **1997**, *77*, 169-195, doi:https://doi.org/10.1016/S0016-7061(97)00021-9.

34. Demattê, J.A.M., Sousa, A.A., Alves, M.C., Nanni, M.R., Fiorio, P.R., Campos, R.C. Determining soil water status and other soil characteristics by spectral proximal sensing. *Geoderma*, **2006**, 135, 179-195.

35. Deng, X.; Wang, Y.; Yun, R.; Peng, H. Clustering of high-resolution remote sensing imagery. In Proceedings of Third International Asia-Pacific Environmental Remote Sensing Remote Sensing of the Atmosphere, Ocean, Environment, and Space, Hangzhou, China, June, 2003.

36. Dharumarajan, S.; Hegde, R.; Singh, S.K. Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India. *Geoderma Regional* **2017**, *10*, 154-162, doi:10.1016/j.geodrs.2017.07.005.

37. Dhawale, N.; Adamchuk, V.; Huang, H.; Ji, W.; Lauzon, S.; Biswas, A.; Dutilleul, P. Integrated Analysis of Multilayer Proximal Soil Sensing Data. In Proceedings of Proceedings of the 13th International Conference on Precision Agriculture, St. Louis, Missouri, USA, 31 July–4 August, 2016.

38. Dhawale, N.M.; Adamchuk, V.I.; Prasher, S.O.; Dutilleul, P.R.L.; Ferguson, R.B. Spatially Constrained Geospatial Data Clustering for Multilayer Sensor-Based Measurements. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2014**, *XL-2*, 187-190, doi:10.5194/isprsarchives-XL-2-187-2014.

39. Dierke, C.; Werban, U. Relationships between gamma-ray data and soil properties at an agricultural test site. *Geoderma* **2013**, *199*, 90-98, doi:10.1016/j.geoderma.2012.10.017.

40. Duda, B.M., Weindorf, D.C., Chakraborty, S., Li, B., Man, T., Paulette, L., Deb, S. Soil characterization across catenas via advanced proximal sensors. *Geoderma*, **2017**, 298, 78-91.

41. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. Density-Based Clustering Algorithms for Discovering Clusters. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, 2–4 August 1996; pp. 226–231.

42. Fisher, D. Iterative Optimization and Simplification of Hierarchical Clustering. *J. Artif. Intell. Res.* **1996**, *4*, 147–178.

43. Fraisse, C.W.; Sudduth, K.A.; Kitchen, N.R. Delineation of Site-Specific Management Zones by Unsupervised Classification. *Trans. ASAE* **2001**, *44*, 155–166.

44. Franzen, D.; Mulla, D.. A History of Precision Agriculture. In *Precision Agriculture Technology for Crop Farming*, Zhang, Q., Ed., Boca Raton, FL, USA: CRC Press, 2015; p. 1–19.

45. Fridgen, J.J.; Kitchen, N.R.; Sudduth, K.A.; Drummond, S.T.; Wiebold, W.J.; Fraisse, C.W. Management zone analyst (MZA): software for subfield management zone delineation. *Agronomy journal* **2004**, *96*, 100.

46. Friedman, S.P. Soil properties influencing apparent electrical conductivity: a review. *Computers and Electronics in Agriculture* **2005**, *46*, 45-70, doi:10.1016/j.compag.2004.11.001.

47. García-Tomillo, A., Mirás-Avalos, J.M., Dafonte-Dafonte, J., Paz-González, A. Estimating soil organic matter using interpolation methods with a electromagnetic induction sensor and topographic parameters: a case study in a humid region. *Precision Agriculture*, **2016**, 18(5), 882-897.

48. Géron, A.l. Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. First edition. ed.; O'Reilly Media: Sebastopol, CA, 2017.

49. Gitelson, A.A.; Viña, A.; Arkebauer, T.J.; Rundquist, D.C.; Keydan, G.; Leavitt, B. Remote estimation of leaf area index and green leaf biomass in maize canopies. *Geophysical Research Letters* **2003**, *30*, doi:10.1029/2002GL016450.

50. GNip, P.G.; Harvát, K.C. Management of Zones in Precision Farming. *Agric. Econ.* **2003**, *49*, 416–418.

51. González-Fernández, A.B.; Rodríguez-Pérez, J.R.; Ablanedo, E.S.; Ordoñez, C. Vineyard Zone Delineation by Cluster Classification Based on Annual Grape and Vine Characteristics. *Precis. Agric.* **2017**, *18*, 525–573.

52. Gregory, S.D.L.; Lauzon, J.D.; O'Halloran, I.P.; Heck, R.J. Predicting soil organic matter content in southwestern Ontario fields using imagery from high-resolution digital cameras. *Canadian Journal of Soil Science* **2006**, *86*, 573-584, doi:10.4141/S05-043.

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

54. Grunwald, S. *Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics (Ed.)*; CRC Press: Boca Raton, 2006; https://doi.org/10.1201/9781420028188.

55. Grunwald, S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma* **2009**, *152*, 195-207, doi:10.1016/j.geoderma.2009.06.003.

56. Grunwald, S.; Thompson, J.A.; Boettinger, J.L. Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues. *Soil Science Society of America Journal* **2011**, *75*, doi:10.2136/sssaj2011.0025.

57. Grunwald, S.; Vasques, G.M.; Rivero, R.G. Fusion of Soil and Remote Sensing Data to Model Soil Properties. In *Advances in Agronomy*, Sparks, D.L., Ed. Academic Press: 2015; Vol. 131, pp. 1-109.

58. Gui-Fen, C.; Li-Ying, C.; Guo-Wei, W.; Bao-Cheng, W.; Da-You, L.; Sheng-Sheng, W. Application of a spatial fuzzy clustering algorithm in precision fertilisation. *New Zealand Journal of Agricultural Research* **2007**, *50*, 1249-1254, doi:10.1080/00288230709510409.

59. Guo, P.T.; Li, M.F.; Luo, W.; Tang, Q.F.; Liu, Z.W.; Lin, Z.M. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. *Geoderma* **2015**, *237-238*, 49-59, doi:10.1016/j.geoderma.2014.08.009.

60. Hartigan, J.A.; Wong, M.A. A K-Means Clustering Algorithm. *Appl. Stat.* **2012**, *28*, 100–108.

61. Hatfield, J.L.; Gitelson, A.A.; Schepers, J.S.; Walthall, C.L. Application of Spectral Remote Sensing for Agronomic Decisions. *Agronomy Journal* **2008**, *100*, 117-131, doi:10.2134/agronj2006.0370c.

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

63. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.; Gonzalez, M.R.; Kilibarda, M.; Blagotic, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B., et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* **2017**, *12*, e0169748, doi:10.1371/journal.pone.0169748.

64. Hengl, T.; Nussbaum, M.; Wright, M.N.; Heuvelink, G.B.M.; Graler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* **2018**, *6*, e5518, doi:10.7717/peerj.5518.

65. Heung, B.; Bulmer, C.E.; Schmidt, M.G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. *Geoderma* **2014**, *214-215*, 141-154, doi:10.1016/j.geoderma.2013.09.016.

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

67. Huang, H.H., Adamchuk, V., Biswas, A., Ji, W., Lauzon, S. Analysis of Soil Properties Predictability Using Different On-the-Go Soil Mapping Systems, In Proceedings of the 14th International Conference on Precision Agriculture, Montréal, Canada, 2018.

68. Huang, J.; Lark, R.M.; Robinson, D.A.; Lebron, I.; Keith, A.M.; Rawlins, B.; Tye, A.; Kuras, O.; Raines, M.; Triantafilis, J. Scope to predict soil properties at within-field scale from small samples using proximally sensed γ-ray spectrometer and EM induction data. *Geoderma* **2014**, *232-234*, 69-80, doi:https://doi.org/10.1016/j.geoderma.2014.04.031.

69. Huang, Y.; Lan, Y.; Thomson, S.J.; Fang, A.; Hoffmann, W.C.; Lacey, R.E. Development of Soft Computing and Applications in Agricultural and Biological Engineering. *Comput. Electron. Agric.* **2010**, *71*, 107–127.

70. Huete, A.R. Remote Sensing for Environmental Monitoring. In *Environmental Monitoring and Characterization*, Artiola, J.F., Pepper, I.L., Brusseau, M.L., Eds. Academic Press: Burlington, 2004; https://doi.org/10.1016/B978-012064477-3/50013-8pp. 183-206.

71. Hurvich, C.M.; Tsai, C. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297-307, doi:10.1093/biomet/76.2.297.

72. Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., Vohland, M. In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. *Geoderma*, **2019**, 355.

73. ISPA. Precision Ag Definition. Availabe online: https://www.ispag.org/ (accessed on **2019**/11/11).

74. Ji, W.; Adamchuk, V.; Lauzon, S.; Su, Y.; Saifuzzaman, M.; Huang, H. Pre-processing of on-the-go mapping data. In Proceedings of Pedometrics 2017 Conference, Wageningen, The Netherlands, 26 June-1 July 2017 p. 113.

75. Ji, W.; Adamchuk, V.I.; Chen, S.; Su, A.S.M.; Ismail, A.; Gan, Q.; Shi, Z.; Biswas, A. Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study. *Geoderma* **2019**, *341*, 111-128, doi:10.1016/j.geoderma.2019.01.006.

76. Jiang, Q.; Fu, Q.; Wang, Z. Study on Delineation of Irrigation Management Zones Based on Management Zone Analyst Software. In *International Conference on Computer and*

*Computing Technologies in Agriculture*; Springer: Berlin/Heidelberg, Germany, 2010; pp 419–427.

77. Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241-254, doi:10.1007/BF02289588.

78. Kaye, J.P.; Majumdar, A.; Gries, C.; Buyantuyev, A.; Grimm, N.B.; Hope, D.; Jenerette, G.D.; Zhu, W.X.; Baker, L. Hierarchical Bayesian Scaling of Soil Properties Across Urban, Agricultural, and Desert Ecosystems. *Ecological Applications* **2008**, *18*, 132-145, doi:10.1890/06-1952.1.

79. Kerry, R.; Goovaerts, P.; Giménez, D.; Oudemans, P.V. Investigating temporal and spatial patterns of cranberry yield in New Jersey fields. *Precision Agriculture* **2017**, *18*, 507-524, doi:10.1007/s11119-016-9471-8.

80. Keskin, H.; Grunwald, S.; Harris, W.G. Digital mapping of soil carbon fractions with machine learning. *Geoderma* **2019**, *339*, 40-58, doi:10.1016/j.geoderma.2018.12.037.

81. Khosla, R., K.; Fleming, K.; Delgado, J.A.; Shaver, T.M.; Westfall, D.G. Use of Site-Specific Management Zones to Improve Nitrogen Management for Precision Agriculture. *J. Soil Water Conserv.* **2002**, *57*, 513–518.

82. Khosla, R.; Westfall, D.G.; Reich, R.M.; Mahal, J.S.; Gangloff, W.J. Spatial Variation and Site-Specific Management Zones. In *Geostatistical Applications for Precision Agriculture*; Oliver, M.A., Ed.; Springer Science: Berlin, Germany, 2010; pp. 195–219.

83. Lazarevic, A.; Xu, X.; Fiez, T.; Obradovic, Z. Clustering-Regression-Ordering Steps for Knowledge Discovery in Spatial Databases. In Proceedings of the International Joint Conference on Neural Networks Washington, DC, USA, 10–16 July 1999; 2530–2534.

84. Li, Y.; Shi, Z.; Li, F. Delineation of Site-Specific Management Zones Based on Temporal and Spatial Variability of Soil Electrical Conductivity. *Pedosphere* **2007**, *17*, 156-164, doi:https://doi.org/10.1016/S1002-0160(07)60021-6.

85. Liu, T.; Abd-Elrahman, A. An Object-Based Image Analysis Method for Enhancing Classification of Land Covers Using Fully Convolutional Networks and Multi-View Images of Small Unmanned Aerial System. *Remote Sensing* **2018**, *10*, 457, doi:https://doi.org/10.3390/rs10030457.

86. Liu, Y.; Xiong, N.; Zhao, Y.; Vasilakos, A.V.; Gao, J.; Jia, Y. Multi-Layer Clustering Routing Algorithm for Wireless Vehicular Sensor Networks. *IET Commun.* **2010**, *4*, 810.

87. Lück, E.; Gebbers, R.; Ruehlmann, J.; Spangenberg, U. Electrical conductivity mapping for precision farming. *Near Surface Geophysics* **2009**, *7*, 15-26, doi:10.3997/1873-0604.2008031.

88. Mahmood, H.S.; Hoogmoed, W.B.; Henten, E.J. Sensor data fusion to predict multiple soil properties. *Precision Agriculture* **2012**, *13*, 628-645, doi:10.1007/s11119-012-9280-7.

89. Malone, B.P.; Jha, S.K.; Minasny, B.; McBratney, A.B. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. *Geoderma* **2016**, *262*, 243-253, doi:10.1016/j.geoderma.2015.08.037.

90. McBratney, A.B.; Pringle, M.J. Estimating Average and Proportional Variograms of Soil Properties and Their Potential Use in Precision Agriculture. *Precision Agriculture* **1999**, *1*, 125-152, doi:10.1023/A:1009995404447.

91. McBratney, A.B.; Santos, M.L.M.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3-52, doi:10.1016/s0016-7061(03)00223-4.

92. McFadden, B.R.; Brorsen, B.W.; Raun, W.R. Nitrogen fertilizer recommendations based on plant sensing and Bayesian updating. *Precision Agriculture* **2017**, *19*, 79-92, doi:10.1007/s11119-017-9499-4.

93. Meier, M.; Souza, E.d.; Francelino, M.R.; Filho, E.I.F.; Schaefer, C.E.G.R. Digital Soil Mapping Using Machine Learning Algorithms in a Tropical Mountainous Area. *Revista Brasileira de Ciência do Solo* **2018**, *42*, 1-22, doi:10.1590/18069657rbcs20170421.

94. Meirvenne, M.; Cleemput, I. Pedometrical techniques for soil texture mapping at a regional scale. In *Environmental soil-landsacpe modeling: geographical information technologies and pedometrics*, Grunwald, S., Ed. CRC Press Boca Raton, USA, 2006; pp. 323-341.

95. Merrill, H.R.; Grunwald, S.; Bliznyuk, N. Semiparametric regression models for spatial prediction and uncertainty quantification of soil attributes. *Stochastic Environmental Research and Risk Assessment* **2017**, *31*, 2691-2703, doi:10.1007/s00477-016-1337-0.

96. Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M. Impact of multi-scale predictor selection for modeling soil properties. *Geoderma*, **2015**, 239-240, 97-106.

97. Minasny, B., McBratney, A.B. Why you don't need to use RPD, Pedometron. International Union of Soil Sciences, Österreich/Austria, 2013, pp. 14-15.

98. Minasny, B.; McBratney, A.B. Digital soil mapping: A brief history and some lessons. *Geoderma* **2016**, *264*, 301-311, doi:10.1016/j.geoderma.2015.07.017.

99. Mondal, P.; Jain, M.; Defries, R.S.; Galford, G.L.; Small, C. Sensitivity of Crop Cover to Climate Variability: Insights from Two Indian Agro-Ecoregions. *J. Environ. Manag.* **2014**, *148*, 21–30.

100. Motwani, M. A Study on Initial Centroids Selection for Partitional Clustering Algorithms. *Adv. Intell. Syst. Comput.* **2019**, *731*, 211–220.

101. Mulder, V.L.; de Bruin, S.; Schaepman, M.E.; Mayr, T.R. The use of remote sensing in soil and terrain mapping -A review. *Geoderma* **2011**, *162*, 1-19, doi:10.1016/j.geoderma.2010.12.018.

102. Mulla, D.J. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering* **2013**, *114*, 358-371, doi:https://doi.org/10.1016/j.biosystemseng.2012.08.009.

103. Nazeer, K.A.A.; Sebastian, M.P. Improving the Accuracy and Efficiency of the k-Means Clustering Algorithm. *Proc. World Cong. Eng.* **2009**, *I*, 1–5.

104. Neely, H.L., Morgan, C.L.S., Hallmark, C.T., McInnes, K.J., Molling, C.C. Apparent electrical conductivity response to spatially variable vertisol properties. *Geoderma*, **2016**, 263, 168-175.

105. Nguy-Robertson, A.; Gitelson, A.; Peng, Y.; Viña, A.; Arkebauer, T.; Rundquist, D. Green Leaf Area Index Estimation in Maize and Soybean: Combining Vegetation Indices to Achieve Maximal Sensitivity. *Agronomy Journal* **2012**, *104*, 1336-1347, doi:10.2134/agronj2012.0065.

106. Nocco, M.A., Ruark, M.D., Kucharik, C.J. Apparent electrical conductivity predicts physical properties of coarse soils. *Geoderma*, **2019**, 335, 1-11.

107. Odeha, I.O.A.; McBratney, A.B.; Chittleborough, D.J. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma* **1994**, *63*, 197-214, doi:https://doi.org/10.1016/0016-7061(94)90063-9.

108. Oliver, M.A. An Overview of Geostatistics and Precision Agriculture. In: *Geostatistical Applications for Precision Agriculture*, Oliver M.A., Ed., Springer Netherlands, Dordrecht, 2010; p. 1-34.

109. Oliver, M.A. Precision agriculture and geostatistics: How to manage agriculture more exactly. *Significance* **2013**, *10*, 17-22, doi:10.1111/j.1740-9713.2013.00646.x.

110. Orhan, U.; Hekim, M.; Ozer, M. EEG Signals Classification Using the *K*-Means Clustering and a Multilayer Perceptron Neural Network Model. *Expert Syst. Appl.* **2011**, *38*, 13475–13481.

111. Padarian, J.; Minasny, B.; McBratney, A.B. Using deep learning for digital soil mapping. *Soil* **2019**, *5*, 79-89, doi:10.5194/soil-5-79-2019.

112. Panayi, E.; Peters, G.W.; Kyriakides, G. Statistical modelling for precision agriculture: A case study in optimal environmental schedules for Agaricus Bisporus production via variable domain functional regression. *PLoS One* **2017**, *12*, e0181921, doi:10.1371/journal.pone.0181921.

113. Panda, S.; Sahu, S.; Jena, P.; Chattopadhyay, S. Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study. In Proceedings of Advances in Computer Science, Engineering & Applications, Berlin, Heidelberg, 2012; pp. 451-460.

114. Pelletier, C.; Valero, S.; Inglada, J.; Champion, N.; Dedieu, G. Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment* **2016**, *187*, 156-168, doi:https://doi.org/10.1016/j.rse.2016.10.010.

115. Pierpaoli, E., Carli, G., Pignatti, E., Canavari, M. Drivers of Precision Agriculture Technologies Adoption: A Literature Review. *Procedia Technology*, **2013**, 8, 61-69.

116. Piikki, K.; Söderström, M.; Stenberg, B. Sensor data fusion for topsoil clay mapping. *Geoderma* **2013**, *199*, 106-116, doi:10.1016/j.geoderma.2012.10.007.

117. Poppiel, R.R.; Lacerda, M.P.C.; Demattê, J.A.M.; Oliveira, M.P.; Gallo, B.C.; Safanelli, J.L. Pedology and soil class mapping from proximal and remote sensed data. *Geoderma* **2019**, *348*, 189-206, doi:10.1016/j.geoderma.2019.04.028.

118. Pouladi, N.; Møller, A.B.; Tabatabai, S.; Greve, M.H. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. *Geoderma* **2019**, *342*, 85-92, doi:10.1016/j.geoderma.2019.02.019.

119. Qi, J.; Chehbouni, A.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A modified soil adjusted vegetation index. *Remote Sensing of Environment* **1994**, *48*, 119-126, doi:https://doi.org/10.1016/0034-4257(94)90134-1.

120. Qi, Y. Random Forest for Bioinformatics. In *Ensemble Machine Learning*, Zhang, C., Ma, Y.Q., Eds. Springer: United States, 2012; http://dx.doi.org/10.1007/978-1-4419-9326-7_11pp. 307-323.

121. Rad, M.R.P.; Toomanian, N.; Khormali, F.; Brungard, C.W.; Komaki, C.B.; Bogaert, P. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. *Geoderma* **2014**, *232-234*, 97-106, doi:10.1016/j.geoderma.2014.04.036.

122. Rasaei, Z.; Bogaert, P. Spatial filtering and Bayesian data fusion for mapping soil properties: A case study combining legacy and remotely sensed data in Iran. *Geoderma* **2019**, *344*, 50-62, doi:10.1016/j.geoderma.2019.02.031.

123. Rizzo, R.; Demattê, J.A.M.; Lepsch, I.F.; Gallo, B.C.; Fongaro, C.T. Digital soil mapping at local scale using a multi-depth Vis–NIR spectral library and terrain attributes. *Geoderma* **2016**, *274*, 18-27, doi:10.1016/j.geoderma.2016.03.019.

124. Roberts, D.F.; Adamchuk, V.I.; Shanahan, J.F.; Ferguson, R.B.; Schepers, J.S. Estimation of Surface Soil Organic Matter Using a Ground-Based Active Sensor and Aerial Imagery. *Precis. Agric.* **2011**, *12*, 82–102.

125. Rodrigues, F.A., Bramley, R.G.V., Gobbett, D.L. Proximal soil sensing for Precision Agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in contrasting soils. *Geoderma*, **2015**, 243-244, 183-195.

126. Rodriguez-Moreno, F.; Kren, J.; Zemek, F.; Novak, J.; Lukas, V.; Pikl, M. Advantage of multispectral imaging with sub-centimeter resolution in precision agriculture: generalization of training for supervised classification. *Precision Agriculture* **2017**, *18*, 615-634, doi:10.1007/s11119-016-9478-1.

127. Rouze, G.S.; Morgan, C.L.S.; McBratney, A.B. Understanding the utility of aerial gamma radiometrics for mapping soil properties through proximal gamma surveys. *Geoderma* **2017**, *289*, 185-195, doi:10.1016/j.geoderma.2016.12.004.

128. Ruß, G.; Brenning, A. Data Mining in Precision Agriculture: Management of Spatial Information. In Proceedings of Computational Intelligence for Knowledge-Based Systems Design, Berlin, Heidelberg, 2010; pp. 350-359.

129. Ruß, G.; Kruse, R. Exploratory Hierarchical Clustering for Management Zone Delineation in Precision Agriculture. In Proceedings of Advances in Data Mining. Applications and Theoretical Aspects, Berlin, Heidelberg, 2011; pp. 161-173.

130. Sadahiro, Y. Cluster Perception in the Distribution of Point Objects. *Cartogr. Int. J. Geogr. Inf. Geovisualization* **1997**, *34*, 49–62.

131. Saey, T., Van Meirvenne, M., Vermeersch, H., Ameloot, N., Cockx, L. A pedotransfer function to evaluate the soil profile textural heterogeneity using proximally sensed apparent electrical conductivity. *Geoderma*, **2009**, 150(3-4), 389-395.

132. Saifuzzaman, M.; Adamchuk, V.; Buelvas, R.; Biswas, A.; Prasher, S.; Rabe, N.; Aspinall, D.; Ji, W. Clustering Tools for Integration of Satellite Remote Sensing Imagery and Proximal Soil Sensing Data. *Remote Sensing* **2019**, *11*, 1036.

133. Saifuzzaman, M.; Adamchuk, V.; Huang, H.-H.; Ji, W.; Rabe, N.; Biswas, A. Data Clustering Tools for Understanding Spatial Heterogeneity in Crop Production by Integrating Proximal Soil Sensing and Remote Sensing Data. In Proceedings of Proceedings of the 14th International Conference on Precision Agriculture, Montreal, QC, Canada, 24–27 June 2018; p. 14.

134. Salama, R.B. Remote Sensing of Soils and Plants Imagery. In *Encyclopedia of Agrophysics*, Gliński, J., Horabik, J., Lipiec, J., Eds. Springer Netherlands: Dordrecht, 2011; 10.1007/978-90-481-3585-1_132pp. 681-693.

135. Samet, H. An Overview of Hierarchical Spatial Data Structures. In Proceedings of the Fifth Israeli Symposium on Artificial Intelligence, Vision, and Pattern Recognition, Tel-Aviv, Ganei-Hata`arucha, Israel, 27-28 December 1988; pp. 331–351.

136. Samuel-Rosa, A.; Heuvelink, G.B.M.; Vasques, G.M.; Anjos, L.H.C. Do more detailed environmental covariates deliver more accurate soil maps? *Geoderma* **2015**, *243-244*, 214-227, doi:10.1016/j.geoderma.2014.12.017.

137. Schueller, J.K. Geostatistics and Precision Agriculture: A Way Forward. In *Geostatistical Applications for Precision Agriculture*, Oliver, M.A., Ed. Springer Netherlands: Dordrecht, 2010; 10.1007/978-90-481-9133-8_12pp. 305-312.

138. Shatar, T.M.; McBratney, A. Subdividing a Field into Contiguous Management Zones Using a K-Zones Algorithm. In Proceedings of Third European Conference on Precision Agriculture, Montpellier, France, June 18-20, 2001; pp. 115–120.

139. Singh, G., Williard, K., Schoonover, J. Spatial Relation of Apparent Soil Electrical Conductivity with Crop Yields and Soil Properties at Different Topographic Positions in a Small Agricultural Watershed. *Agronomy*, **2016**, 6(4).

140. Söderström, M.; Sohlenius, G.; Rodhe, L.; Piikki, K. Adaptation of regional digital soil mapping for precision agriculture. *Precision Agriculture* **2016**, *17*, 588-607, doi:10.1007/s11119-016-9439-8.

141. Sommer, M.; Wehrhan, M.; Zipprich, M.; Weller, U.; zu Castell, W.; Ehrich, S.; Tandler, B.; Selige, T. Hierarchical data fusion for mapping soil units at field scale. *Geoderma* **2003**, *112*, 179-196, doi:https://doi.org/10.1016/S0016-7061(02)00305-1.

142. Stockmann, U., Huang, J., Minasny, B., Triantafilis, J., Goss, M. Utilizing a DUALEM-421 and inversion modelling to map baseline soil salinity along toposequences in the Hunter Valley Wine district. *Soil Use and Management*, **2017**, 33(3), 413-424.

143. Sudduth, K.A.; Myers, D.B.; Kitchen, N.R.; Drummond, S.T. Modeling soil electrical conductivity–depth relationships with data from proximal and penetrating ECa sensors. *Geoderma* **2013**, *199*, 12-21, doi:https://doi.org/10.1016/j.geoderma.2012.10.006.

144. Sun, Y.; Druecker, H.; Hartung, E.; Hueging, H.; Cheng, Q.; Zeng, Q.; Sheng, W.; Lin, J.; Roller, O.; Paetzold, S., et al. Map-based investigation of soil physical conditions and crop

yield using diverse sensor techniques. *Soil & Tillage Research* **2011**, *112*, 149-158, doi:10.1016/j.still.2010.12.002.

145. Szatmári, G.; Pásztor, L. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. *Geoderma* **2019**, *337*, 1329-1340, doi:10.1016/j.geoderma.2018.09.008.

146. Taylor, J.C.; Wood, G.A.; Earl, R.; Godwin, R.J. Soil Factors and their Influence on Within-field Crop Variability, Part II: Spatial Analysis and Determination of Management Zones. *Biosystems Engineering* **2003**, *84*, 441-453, doi:10.1016/s1537-5110(03)00005-9.

147. U.S. Department of Agriculture (USDA). *Management Zone Analyst Version 1.0 Software*; U.S. Department of Agriculture: Washington, DC, USA, 2000.

148. Vaysse, K.; Lagacherie, P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* **2017**, *291*, 55-64, doi:10.1016/j.geoderma.2016.12.017.

149. Vendrusculo, L.G.; Kaleita, A.F. Modeling zone management in precision agriculture through Fuzzy C-Means technique at spatial database. In Proceedings of Agricultural and Biosystems Engineering Conference Louisville, KY, USA, August 7-10, 2011; pp. 2701–2715.

150. Vermeulen, D.; Niekerk, A.V. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. *Geoderma* **2017**, *299*, 1-12, doi:10.1016/j.geoderma.2017.03.013.

151. Veum, K.S.; Sudduth, K.A.; Kremer, R.J.; Kitchen, N.R. Sensor data fusion for soil health assessment. *Geoderma* **2017**, *305*, 53-61, doi:10.1016/j.geoderma.2017.05.031.

152. Viña, A.; Gitelson, A.A.; Nguy-Robertson, A.L.; Peng, Y. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. *Remote Sensing of Environment* **2011**, *115*, 3468-3478, doi:https://doi.org/10.1016/j.rse.2011.08.010.

153. Viscarra Rossel, R.A., Brus, D.J., Lobsey, C., Shi, Z., McLachlan, G. Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference. *Geoderma*, **2016**, 265, 152-163.

154. Viscarra Rossel, R.A.; Adamchuk, V.I. Proximal Soil Sensing. In *Precision Agriculture for Sustainability and Environmental Protection*, Marchant, B., Oliver, M., Bishop, T., Eds. Earthscan from Routledge: New York, 2013; pp. 99–118.

155. Viscarra Rossel, R.A.; Adamchuk, V.I.; Sudduth, K.A.; McKenzie, N.J.; Lobsey, C. Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. In *Advances in Agronomy*, Sparks, D.L., Ed. Academic Press: 2011; Vol. 113, pp. 243-291.

156. Vitharana, U.W.A., Saey, T., Cockx, L., Simpson, D., Vermeersch, H., Van Meirvenne, M. Upgrading a 1/20,000 soil map with an apparent electrical conductivity survey. *Geoderma*, **2008**, 148(1), 107-112.

157. Vrindts, E.; Mouazen, A.M.; Reyniers, M.; Maertens, K.; Maleki, M.R.; Ramon, H.; De Baerdemaeker, J. Management Zones based on Correlation between Soil Compaction, Yield and Crop Data. *Biosystems Engineering* **2005**, *92*, 419-428, doi:10.1016/j.biosystemseng.2005.08.010.

158. Wadoux, A.M.J.C. Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma* **2019**, *351*, 59-70, doi:10.1016/j.geoderma.2019.05.012.

159. Walker, E., Monestiez, P., Gomez, C., Lagacherie, P. Combining measured sites, soilscapes map and soil sensing for mapping soil properties of a region. *Geoderma*, **2017**, 300, 64-73.

160. Walters, R.W.; Jenq, R.R.; Hall, S.B. Evaluating Farmer Defined Management Zone Maps for Variable Rate Fertilizer Application. *Precis. Agric.* **2000**, *2*, 201–215.

161. Watson, H.D., Neely, H.L., Morgan, C.L.S., McInnes, K.J., Molling, C.C. Identifying subsoil variation associated with gilgai using electromagnetic induction. *Geoderma*, **2017**, 295, 34-40.

162. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. Deep learning. In *Data Mining: Practical Machine Learning Tools and Techniques* 4th ed.; Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., Eds. Morgan Kaufmann: United States, 2017; https://doi.org/10.1016/B978-0-12-804291-5.00010-6pp. 417-466.

163. Wulf, H.; Mulder, T.; Schaepman, M.; Keller, A.; Jorg, P.C. *Remote sensing of soils*; University of Zurich: Zürich 2015/01/22, 2015.

164. Xiong, X.; Grunwald, S.; Myers, D.B.; Kim, J.; Harris, W.G.; Bliznyuk, N. Assessing uncertainty in soil organic carbon modeling across a highly heterogeneous landscape. *Geoderma* **2015**, *251-252*, 105-116, doi:https://doi.org/10.1016/j.geoderma.2015.03.028.

165. Xu, Y.; Smith, S.E.; Grunwald, S.; Abd-Elrahman, A.; Wani, S.P.; Nair, V.D. Estimating soil total nitrogen in smallholder farm settings using remote sensing spectral indices and regression kriging. *Catena* **2018**, *163*, 111-122, doi:10.1016/j.catena.2017.12.011.

166. Xue, J.; Su, B. Significant Remote Sensing Vegetation Indices : A Review of Developments and Applications. *Journal of Sensors* **2017**, *2017*, 17, doi:https://doi.org/10.1155/2017/1353691.

167. Yan, L.; Zhou, S.; Feng, L. Delineation of Site-Specific Management Zones Based on Temporal and Spatial Variability of Soil Electrical Conductivity. *Pedosphere* **2007**, *17*, 156–164.

168. Zare, E.; Beucher, A.; Huang, J.; Boman, A.; Mattback, S.; Greve, M.H.; Triantafilis, J. Three-dimensional imaging of active acid sulfate soil using a DUALEM-21S and EM inversion software. *Journal of Environmental Management* **2018**, *212*, 99-107, doi:10.1016/j.jenvman.2018.02.008.

169. Zeraatpisheh, M.; Ayoubi, S.; Jafari, A.; Tajik, S.; Finke, P. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. *Geoderma* **2019**, *338*, 445-452, doi:10.1016/j.geoderma.2018.09.006.

170. Zhang, N.; Wang, M.; Wang, N. Precision agriculture-a worldwide overview. *Computers and Electronics in Agriculture* **2002**, *36*, 113-132, doi:https://doi.org/10.1016/S0168-1699(02)00096-0.

171. Zhou, J.; Khot, L.R.; Bahlol, H.Y.; Boydston, R.; Miklas, P.N. Evaluation of ground, proximal and aerial remote sensing technologies for crop stress monitoring. *IFAC-PapersOnLine* **2016**, *49*, 22-26, doi:https://doi.org/10.1016/j.ifacol.2016.10.005.

172. Zhou, J.; Li, E.; Wei, H.; Li, C.; Qiao, Q.; Armaghani, D.J. Random Forests and Cubist Algorithms for Predicting Shear Strengths of Rockfill Materials. *Applied Sciences* **2019**, *9*, 1621, doi:10.3390/app9081621.

# Appendices

## A. Data portal all study sites

**Table A1** PSS and soil sample data web portal: All field data collected from Ontario and preserved in web repository for this research.

| Experimental fields | Field code | Available data | Chapter numbers |
|---|---|---|---|
| ON_Hunter_GFOND | WH | Field boundary<br>Field elevation (RTK) - 2014 & 2016<br>DUALEM-21S ($EC_a$) sensing - 2016<br>Laboratory analysis (soil sample points: 99) - 2014 | 3 & 4 |
| ON_Linders_GFO | LD | Field boundary<br>Field elevation (RTK) - 2015<br>DUALEM-21S (ECa) sensing - 2015<br>Gamma-Ray (SoilOptix) sensing - 2015<br>Laboratory analysis (soil sample points: 62) - 2015 | 3, 4, & 5 |
| ON_Rainbarrel_GFO | RB | Field boundary<br>Field elevation (RTK) - 2015<br>DUALEM-21S ($EC_a$) sensing - 2015<br>Laboratory analysis (soil sample points: 72) - 2014 | 3 & 4 |
| ON_Field25_GFO | F25 | Field boundary<br>Field elevation (RTK) - 2014<br>DUALEM-21S ($EC_a$) sensing - 2015<br>Laboratory analysis (soil sample points: 26) -2014 | 4 |
| ON_Kenmore_GFO | KM | Field boundary<br>Field elevation (RTK) - 2016<br>DUALEM-21S ($EC_a$) sensing - 2016<br>Laboratory analysis (soil sample points: 119) - 2014 | 4 |
| ON_Lamport_GFO | LP | Field boundary<br>Field elevation (RTK) – 2012 & 2015<br>DUALEM-21S ($EC_a$) sensing - 2015<br>Laboratory analysis (soil sample points: 72) - 2014 | 4 |
| ON_Line_GFOND | TE | Field boundary<br>Field elevation (RTK) - 2014<br>DUALEM-21S ($EC_a$) sensing - 2015<br>Laboratory analysis (soil sample points: 97) - 2014 | 4 |
| ON_McCarter_GFO | SM | Field boundary<br>Field elevation (RTK) - 2015<br>DUALEM-21S ($EC_a$) sensing - 2015<br>Laboratory analysis (soil sample points: 74) - 2014 | 4 |
| ON_Nixon_ND | NX | Field boundary<br>Field elevation (RTK) - 2017<br>DUALEM-21S ($EC_a$) sensing - 2017<br>Laboratory analysis (soil sample points: 74)- 2015 & 2017 | 4 |

| | | | |
|---|---|---|---|
| ON_R50_GFO | R50 | Field boundary<br>Field elevation (RTK) - 2014<br>DUALEM-21S ($EC_a$) sensing - 2015<br>Laboratory analysis (soil sample points: 51) - 2014 | 4 |
| ON_Rhineland_GFO | RL | Field boundary<br>Field elevation (RTK) - 2015<br>DUALEM-21S ($EC_a$) sensing - 2015<br>Laboratory analysis (soil sample points: 49) - 2014 | 4 |
| ON_Schouten_ND | ST | Field boundary<br>Field elevation (RTK) - 2016<br>DUALEM-21S ($EC_a$) sensing - 2016<br>Laboratory analysis (soil sample points: 76) - 2016 | 4 |
| ON_Vernon_GFO | VN | Field boundary<br>Field elevation (RTK) - 2016<br>DUALEM-21S ($EC_a$) sensing - 2016<br>Laboratory analysis (soil sample points: 51) - 2014 | 4 |

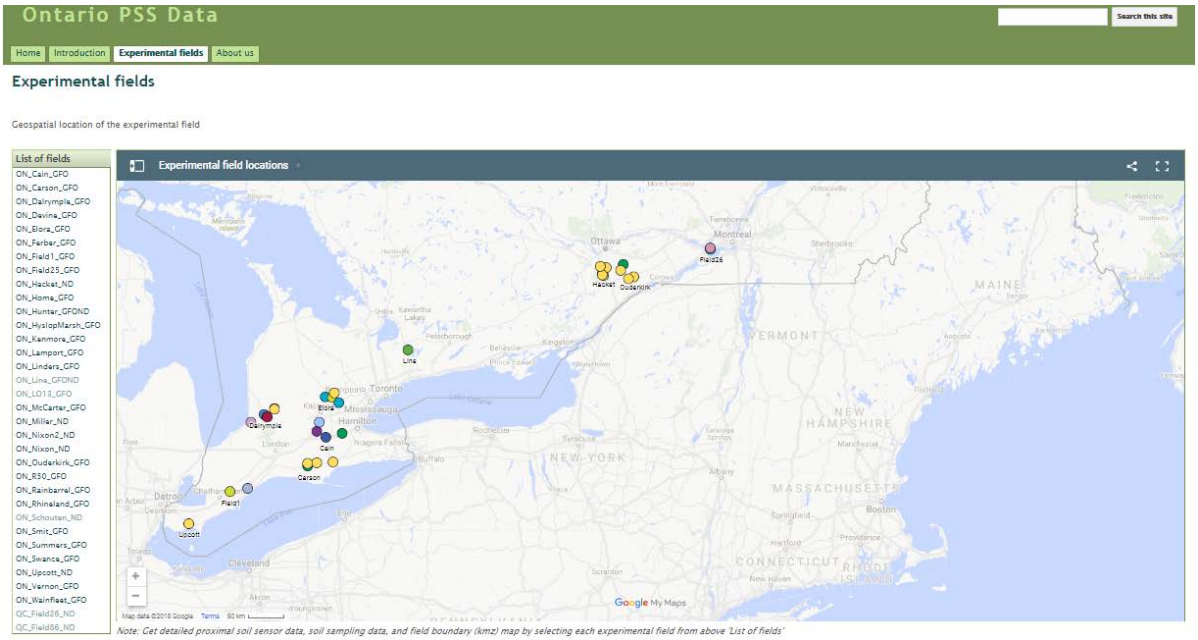



**Figure A1** Web interface for the data repository: https://sites.google.com/site/omaframcgill2016/project-sites

## B. Scripts for clustering software in Chapter 3

### B1 - Python scripts for NSA clustering

```python
# md.saifuzzaman@mail.mcgill.ca

# Data import
#############################################################################################################
import math
import pickle
import numpy as np
import scipy.signal
import pandas as pd
from tkinter import *
from tkinter import filedialog
import matplotlib.pyplot as pyplot

#np.set_printoptions(threshold=np.inf)

def latconv(lat,minc,F_lat):
   return (lat-minc)*F_lat

def longconv(long,mind,F_long):
   return (long-mind)*F_long

#####################################
#Step 1:Import data from text and save
#####################################

root = Tk()
#root.withdraw()
filename = filedialog.askopenfilename(filetypes = (("Template files", "*.txt"), ("All files", "*")))
root.destroy()
if len(filename) > 0:
    print("You chose %s" % filename)

#filename='dualem'
nsaData = pd.read_table(filename, sep='\t', header='infer', names=None, index_col=False, usecols=None)
labels=nsaData.columns.values
#####################################
#Step 2:Plane coordinates conversion
#####################################

if 'Elevation' in nsaData.columns:
    h=nsaData['Elevation'].mean()#height over elipsoid
else:
    h=200 # Where the data has no elevation information
a=6378137 #semimajor axis
b=6356752.3142 #semiminor axis
c=nsaData['Latitude'].mean()#Avarage latitude
d=math.sqrt((a*math.cos(c))**2+(b*math.sin(c))**2)
F_long=(np.pi*math.cos(c)/180)*((a**2/d)+h) # Longitude factor
F_lat=(np.pi/180)*(((a*b)**2/d**3)+h) # Latitude factor

LongMin = nsaData['Longitude'].min()
LongMax = nsaData['Longitude'].max()
LatMin = nsaData['Latitude'].min()
LatMax = nsaData['Latitude'].max()
convParams = [F_long,LongMin,F_lat,LatMin]

#planer matrix Lat
nsaData['Lat_y'] = nsaData['Latitude'].apply(lambda row : latconv(row,LatMin,F_lat))

#Planer matrix Long
nsaData['Long_x'] = nsaData['Longitude'].apply(lambda row : longconv(row,LongMin,F_long))

```

```
#######################################
#Step 3:Create square grid and cell size
#######################################

gdsz = 40 # cell size
gdc = gdsz / 2 # center of the grid cell
X = np.array(nsaData['Long_x'])
Y = np.array(nsaData['Lat_y'])
Z = np.array(nsaData.iloc[:,3:-2])
xmin=X.min()
xmax=X.max()
ymin=Y.min()
ymax=Y.max()
Xr = np.linspace(xmin, xmax, int((xmax - xmin)/gdsz))
Xc=Xr[0:-1]+gdsz/2
Yr = np.linspace(ymin, ymax, int((ymax - ymin)/gdsz))
```

```
Yc=Yr[0:-1]+gdsz/2
ngx=len(Xc)
ngy=len(Yc)
[nd,nv]=Z.shape
ar=np.zeros((nv, ngy, ngx))
zar=np.zeros((ngy,ngx),dtype=int)
for l in range(ngx):
    for m in range(ngy):
        for n in range(nd):
            if max(abs(X[n]-Xc[l]),abs(Y[n]-Yc[m]))<=gdsz/2:
                zar[m,l]+=1
for l in range(ngx):
    for m in range(ngy):
        i = 0
        aux = np.zeros((zar[m, l],nv))
        for n in range(nd):
            if max(abs(X[n]-Xc[l]),abs(Y[n]-Yc[m]))<=gdsz/2:
                aux[i, :] = Z[n, :]
                i += 1
        if(zar[m, l]==0): # Masking by field area by outside 0
            ar[:, m, l]=np.zeros((1,nv))
        else:
            ar[:, m, l]=np.mean(aux, axis=0)
for l in range(ngx):
    for m in range(ngy):
        if(zar[m,l]>0):
            zar[m, l]=1

```

```
#######################################
#Step 4: Median filtering, plot and save
#######################################

for i in range(nv):
    ar[i,:,:]=scipy.signal.medfilt2d(ar[i,:,:], 5) # 5 x 5 median filtering
    arMasked=np.ma.masked_where(ar[i,:,:]==0, ar[i,:,:])
    pyplot.figure()
    pyplot.imshow(arMasked, interpolation='none', origin='lower',extent=[0,ngx*gdsz,0,ngy*gdsz])
    pyplot.title(labels[i+3])
    pyplot.ylabel('Northing')
    pyplot.xlabel('Easting')
    pyplot.colorbar()
    ax=pyplot.gca()
    ax.set_xticks(np.arange(gdsz, gdsz*(ngx+1), gdsz), minor=True)
    ax.set_yticks(np.arange(gdsz, gdsz*(ngy+1), gdsz), minor=True)
    ax.grid(which='minor', color='k', linestyle='-', linewidth=1)
    pyplot.savefig(labels[i+3]+'.png', dpi=200)
    pyplot.close()


z = ar.shape
#print (zar.shape)
#print (ar.shape)
with open('NSATemp.pickle', 'wb') as outfile:
    pickle.dump([zar,ar,z,gdsz,labels,convParams],outfile)
```

```
133
134 # NSA functions
135 ##########################################################################################################
136
137 import pickle
138 import numpy as np
139 import xlsxwriter as xl
140 import matplotlib.pyplot as pyplot
141
142 np.set_printoptions(threshold=np.nan)
143
144 def mse(valuesmatrix, groupmatrix, mm, NN):
145     acum=0
146     emptygroups=0
147     for j in range(mm):
148         aux=valuesmatrix[groupmatrix==j+1]
149         nj=aux.size
150         if(nj>0):
151             acum+=aux.var()*nj
152         else:
153             emptygroups+=1
154     return acum/(NN+emptygroups-mm) # MSE calculation
155
156 def validLocations(groupmatrix,nx,ny):
157     validmatrix=np.zeros((ny,nx),dtype=bool)
158     ww=0
159     for j in range(nx-2):
160         for i in range(ny-2):
161             valid=np.prod(groupmatrix[i:i+3,j:j+3])
162             if valid>0:
163                 validmatrix[i,j]=True
164                 ww+=1
165     return validmatrix,ww
166
167 def sdvFunc(valuesmatrix, validmatrix, ww, nx, ny):
168     acum=0
169     for j in range(nx-2):
170         for i in range(ny-2):
171             if validmatrix[i,j]:
172                 acum+=valuesmatrix[i:i+3,j:j+3].var()
173     if ww==0:
174         return 0
175     else:
176         return acum/ww
177
178 # R2 value will be maximum(1) when MSE is minimum
179
180 def calculateOF(valuesmatrix, groupmatrix, mm, NN, ffdv, rr2, rr2max):
181     of=1
182     for i in range(nv):
183         rr2[i]=1.0-mse(valuesmatrix[i,:,:],groupmatrix,mm,NN)/ffdv[i]
184         of*=rr2[i]**rr2max[i]
185     return of
186     #return np.prod(np.power(rr2,rr2max))
187
188 def addGroup(valuesmatrix, groupmatrix, validmatrix, mm, NN, ffdv, rr2, rr2max, nx, ny):
189     mof=0
190     iflag=False
191     for j in range(nx-2):
192         for i in range(ny-2):
193             if(validmatrix[i+1,j+1] and np.prod(groupmatrix[i:i+3,j:j+3]==1)==1):
194                 aux=groupmatrix[i:i+3,j:j+3].copy()
195                 groupmatrix[i:i+3,j:j+3]=(mm+1)*np.ones((3,3))
196                 cof=calculateOF(valuesmatrix, groupmatrix, mm+1, NN, ffdv, rr2, rr2max)
197                 groupmatrix[i:i+3,j:j+3]=aux
198                 if cof>=mof:
199                     iflag=True
200                     mof=cof
201                     mi=i+1
202                     mj=j+1
```

```
    if(iflag):
        return [mof, mi, mj]
    else:
        return [mof]

def extendGroup(valuesmatrix, groupmatrix, coordinates, mm, NN, ffdv, rr2, rr2max, nx, ny):
    mof=0
    iflag=False
    for k in range(mm):
        for ind in range(len(coordinates[0])):                #if(groupmatrix[i+1,j+1]==1):
            if(groupmatrix[coordinates[0][ind]+2,coordinates[1][ind]+1]==k+1
               or groupmatrix[coordinates[0][ind],coordinates[1][ind]+1]==k+1
               or groupmatrix[coordinates[0][ind]+1,coordinates[1][ind]+2]==k+1
               or groupmatrix[coordinates[0][ind]+1,coordinates[1][ind]]==k+1):
                aux=groupmatrix[coordinates[0][ind]+1,coordinates[1][ind]+1]
                groupmatrix[coordinates[0][ind]+1,coordinates[1][ind]+1]=k+1
                cof=calculateOF(valuesmatrix, groupmatrix, mm, NN, ffdv, rr2, rr2max)
                groupmatrix[coordinates[0][ind]+1,coordinates[1][ind]+1]=aux
                if cof>=mof:
                    iflag=True
                    mof=cof
                    mi=coordinates[0][ind]+1
                    mj=coordinates[1][ind]+1
                    mk=k+1
    if iflag:
        return [mof, mi, mj, mk]
    else:
        return [mof]

##################################################
#Step 5: Data clustering
##################################################
with open('NSATemp.pickle','rb') as infile:
    zar,ar,z,gdsz,labels,convParams=pickle.load(infile)
nv=z[0]
ngy=z[1]
ngx=z[2]
N=zar[zar!=0].size
m=1
oldof=[]
[var,w]=validLocations(zar, ngx, ngy)
fdv=-1*np.ones(nv)
sdv=-1*np.ones(nv)
r2max=-1*np.ones(nv)
for i in range(nv):
        fdv[i]=mse(ar[i,:,:],zar,m,N)
        sdv[i]=sdvFunc(ar[i,:,:],var,w,ngx,ngy)
        r2max[i]=1.0-(sdv[i]/fdv[i])
r2=np.zeros(nv)
oldof.append(calculateOF(ar, zar, m, N, fdv, r2, r2max))
u=addGroup(ar, zar, var, m, N, fdv, r2, r2max, ngx, ngy)
if(u[0]>oldof[-1]):
    m+=1
    zar[u[1]:u[1]+3,u[2]:u[2]+3]=m*np.ones((3,3))
    oldof.append(u[0])
flag=True
while(flag):
    car=np.where(zar[1:-1,1:-1]==1)
    u=addGroup(ar, zar, var, m, N, fdv, r2, r2max, ngx, ngy)
    v=extendGroup(ar, zar, car, m, N, fdv, r2, r2max, ngx, ngy)
    if((u[0]-oldof[-1])>9*(v[0]-oldof[-1])):
        if(u[0]>oldof[-1]):
            m+=1
            zar[u[1]:u[1]+3,u[2]:u[2]+3]=m*np.ones((3,3))
            oldof.append(u[0])
        else:
            flag=False
    elif(v[0]>=oldof[-1]):
        zar[v[1],v[2]]=v[3]
        oldof.append(v[0])
```

```
    else:
        flag=False
    #print(oldof[-1])

################################################
#Step 6: Output formatting
################################################

zarMasked=np.ma.masked_where(zar==0, zar)
pyplot.imshow(zarMasked, interpolation='none', origin='lower',extent=[0,ngx*gdsz,0,ngy*gdsz])
pyplot.title('Zones')
pyplot.ylabel('Northing')
pyplot.xlabel('Easting')
pyplot.colorbar()
ax=pyplot.gca()
ax.set_xticks(np.arange(gdsz, gdsz*(ngx+1), gdsz), minor=True)
ax.set_yticks(np.arange(gdsz, gdsz*(ngy+1), gdsz), minor=True)
ax.grid(which='minor', color='k', linestyle='-', linewidth=1)
pyplot.savefig('zones.png', dpi=200)
pyplot.close()

pyplot.figure()
pyplot.plot(np.power(oldof,1.0/nv))
pyplot.title('Objective function')
pyplot.ylabel('R^2')
pyplot.xlabel('Iteration')
pyplot.savefig('OF.png', dpi=200)
pyplot.close()

with open('zones.txt', 'w') as outfile:
    zar.tofile(outfile,sep=" ", format="%.5f")

with open('result.pickle', 'wb') as outfile:
    pickle.dump([zar,oldof],outfile)

```

```
# Worksheet from the result
workbook = xl.Workbook('result.xlsx')
for i in range(nv):
    worksheet=workbook.add_worksheet(labels[i+3])
    for j in range(ngy):
        worksheet.write_row('A'+str(ngy-j),ar[i,j,:])
worksheet=workbook.add_worksheet('zones')
for j in range(ngy):
    worksheet.write_row('A'+str(ngy-j),zar[j,:])
worksheet=workbook.add_worksheet('stats')
headers=['Minimum','Median','Average','Max','Range']
worksheet.write_row('B1',headers)
ar[ar==0]=np.nan
for i in range(nv):
    worksheet.write_string('A'+str(i+2),labels[i+3])
    worksheet.write_number('B'+str(i+2),np.nanmin(ar[i,:,:]))
    worksheet.write_number('C'+str(i+2),np.nanmedian(ar[i,:,:]))
    worksheet.write_number('D'+str(i+2),np.nanmean(ar[i,:,:]))
    worksheet.write_number('E'+str(i+2),np.nanmax(ar[i,:,:]))
    worksheet.write_formula('F'+str(i+2),'=E'+str(i+2)+'-B'+str(i+2))
worksheet=workbook.add_worksheet('coordinates')
headers2=['Longitude','Latitude','Zone']
worksheet.write_row('A1',headers2)
for j in range(ngx):
    for i in range(ngy):
        worksheet.write_number('A'+str(j*ngy+i+2),gdsz*j/convParams[0]+convParams[1])
        worksheet.write_number('B'+str(j*ngy+i+2),gdsz*i/convParams[2]+convParams[3])
        worksheet.write_number('C'+str(j*ngy+i+2),zar[i,j])
workbook.close()
```

## B2 - Python scripts for k-means clustering

```python
# md.saifuzzaman@mail.mcgill.ca

# k-means
##############################################################################################

from sklearn.cluster import KMeans
from scipy import stats
from scipy import signal
import numpy as np
import math
import pandas as pd
import matplotlib.pyplot as pyplot

#np.set_printoptions(threshold=np.inf)

def latconv(lat,minc,F_lat):
   return (lat-minc)*F_lat

def longconv(long,mind,F_long):
   return (long-mind)*F_long

def sdvFunc(valuesmatrix, validmatrix, ww, nx, ny):
    acum=0.0
    for j in range(nx-2):
        for i in range(ny-2):
            if validmatrix[i+1,j+1]:
                acum+=valuesmatrix[i:i+3,j:j+3].var()
    if ww==0:
        return 0
    else:
        return acum/float(ww)

def validLocations(groupmatrix,nx,ny):
    validmatrix=np.zeros((ny,nx),dtype=bool)
    ww=0
    for j in range(nx-2):
        for i in range(ny-2):
            valid=np.prod(groupmatrix[i:i+3,j:j+3])
            if valid>0:
                validmatrix[i+1,j+1]=True
                ww+=1
    return validmatrix,ww
```

```
43
44  def calculateOF(valuesmatrix, groupmatrix, mm, NN, nnv, ffdv, rr2, rr2max):
45      of=1
46      aux=mse(valuesmatrix,groupmatrix,mm,NN,nnv)
47      for i in range(nnv):
48          rr2[i]=1.0-aux[i]/ffdv[i]
49          if rr2[i]<0:
50              rr2[i]=0
51          of*=rr2[i]**rr2max[i]
52      return of[0]
53
54  def mse(valuesmatrix, groupmatrix, mm, NN, nnv):
55      acum=np.zeros(nnv)
56      for j in range(mm):
57          ind=np.nonzero(groupmatrix==j+1)
58          for k in range(nnv):
59              emptygroups=0
60              aux=valuesmatrix[k,ind[0],ind[1]]
61              nk=aux.size
62              if(nk>0):
63                  acum[k]+=aux.var()*nk
64              else:
65                  emptygroups+=1
66      return acum/(NN+emptygroups-mm)
67
68  ################################
69  # Read file
70  ################################
71  filename = 'Hunter.txt'
72  nsaData = pd.read_table(filename, sep='\t', header='infer', names=None, index_col=False, usecols=None)
73  labels=nsaData.columns.values
74  useTime=False
75  if(useTime):
76      startingColumn=2
77  else:
78      startingColumn=3
79
```

```
80  ################################
81  # Project to planar coordinates
82  ################################
83  if 'Elevation' in nsaData.columns:
84      h=nsaData['Elevation'].mean()#height over elipsoid
85  else:
86      h=200 # Where the data has no elevation information
87  a=6378137 #semimajor axis
88  b=6356752.3142 #semiminor axis
89  c=nsaData['Latitude'].mean()#Avarage latitude
90  d=math.sqrt((a*math.cos(c))**2+(b*math.sin(c))**2)
91  F_long=(np.pi*math.cos(c)/180)*((a**2/d)+h) # Longitude factor
92  F_lat=(np.pi/180)*(((a*b)**2/d**3)+h) # Latitude factor
93  LongMin = nsaData['Longitude'].min()
94  LongMax = nsaData['Longitude'].max()
95  LatMin = nsaData['Latitude'].min()
96  LatMax = nsaData['Latitude'].max()
97  convParams = [F_long,LongMin,F_lat,LatMin]
98  nsaData['Lat_y'] = nsaData['Latitude'].apply(lambda row : latconv(row,LatMin,F_lat))
99  nsaData['Long_x'] = nsaData['Longitude'].apply(lambda row : longconv(row,LongMin,F_long))
100
101 ################################
102 # Apply kMeans
103 ################################
104 numZones=28
105 Z = np.array(nsaData.iloc[:,startingColumn:-2])
106 kmeans = KMeans(n_clusters=numZones).fit(Z)
107
```

```
###############################
# Convert to raster
###############################
gdsz = 20 # cell size
gdc = gdsz / 2 # center of the grid cell
X = np.array(nsaData['Long_x'])
Y = np.array(nsaData['Lat_y'])
kMeansZ = np.array(kmeans.labels_)+1
xmin=X.min()
xmax=X.max()
ymin=Y.min()
ymax=Y.max()
```

```
Xr = np.linspace(xmin, xmax, int((xmax - xmin)/gdsz))
Xc=Xr[0:-1]+gdsz/2
Yr = np.linspace(ymin, ymax, int((ymax - ymin)/gdsz))
Yc=Yr[0:-1]+gdsz/2
ngx=len(Xc)
ngy=len(Yc)
[nd,nv]=Z.shape
ar=np.zeros((nv, ngy, ngx))
zar=np.zeros((ngy,ngx),dtype=int)
kMeansZar=np.zeros((ngy,ngx),dtype=np.uint8)
for l in range(ngx):
    for m in range(ngy):
        for n in range(nd):
            if max(abs(X[n]-Xc[l]),abs(Y[n]-Yc[m]))<=gdsz/2:
                zar[m,l]+=1
for l in range(ngx):
    for m in range(ngy):
        i = 0
        aux = np.zeros((zar[m, l],nv))
        kMeansAux = np.zeros(zar[m,l])
        for n in range(nd):
            if max(abs(X[n]-Xc[l]),abs(Y[n]-Yc[m]))<=gdsz/2:
                aux[i, :] = Z[n, :]
                kMeansAux[i] = kMeansZ[n]
                i += 1
        if(zar[m, l]==0):
            ar[:, m, l]=np.zeros((1,nv))
            kMeansZar[m, l]=0
        else:
            ar[:, m, l]=np.mean(aux, axis=0)
            kMeansZar[m, l]=stats.mode(kMeansAux)[0][0]
for l in range(ngx):
    for m in range(ngy):
        if(zar[m,l]>0):
            zar[m, l]=1

```

```
################################
# Apply median filter
################################
kMeansZar=signal.medfilt2d(kMeansZar, 5)
for i in range(nv):
    ar[i,:,:]=signal.medfilt2d(ar[i,:,:], 5)

################################
# Compute R2
################################
N=zar[zar!=0].size
[var,w]=validLocations(zar, ngx, ngy)
fdv=mse(ar,zar,1,N,nv)
sdv=-1*np.ones(nv)
r2max=-1*np.ones(nv)
for i in range(nv):
    sdv[i]=sdvFunc(ar[i,:,:],var,w,ngx,ngy)
    r2max[i]=1.0-(sdv[i]/fdv[i])
r2=np.zeros((nv,1))
of=calculateOF(ar, kMeansZar, kMeansZar.max(), N, nv, fdv, r2, r2max)
print(np.power(of,1.0/nv))

################################
# Create plots
################################
kMeansZarMasked=np.ma.masked_where(kMeansZar==0, kMeansZar)
pyplot.figure()
pyplot.imshow(kMeansZarMasked, interpolation='none', origin='lower',extent=[0,ngx*gdsz,0,ngy*gdsz])
pyplot.title('kMeans zones')
pyplot.ylabel('Northing')
pyplot.xlabel('Easting')
pyplot.colorbar()
ax=pyplot.gca()
ax.set_xticks(np.arange(gdsz, gdsz*(ngx+1), gdsz), minor=True)
ax.set_yticks(np.arange(gdsz, gdsz*(ngy+1), gdsz), minor=True)
ax.grid(which='minor', color='k', linestyle='-', linewidth=0.1)
pyplot.savefig('kmeans.png', dpi=200)
pyplot.close()
```

**B3 - k-means clustering maps:** Many k-means (5, 15, and 25) clustering maps were produced for preparing Figure 3.12. Those maps were not provided in Chapter 3.

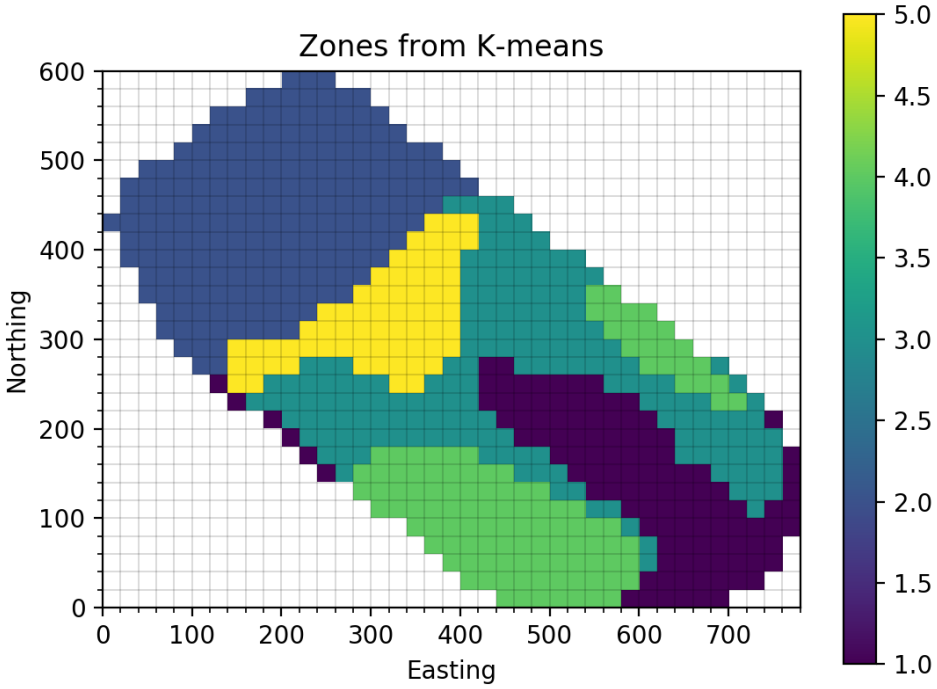



**Figure B1** k-means data clustering for LD field (k=5)

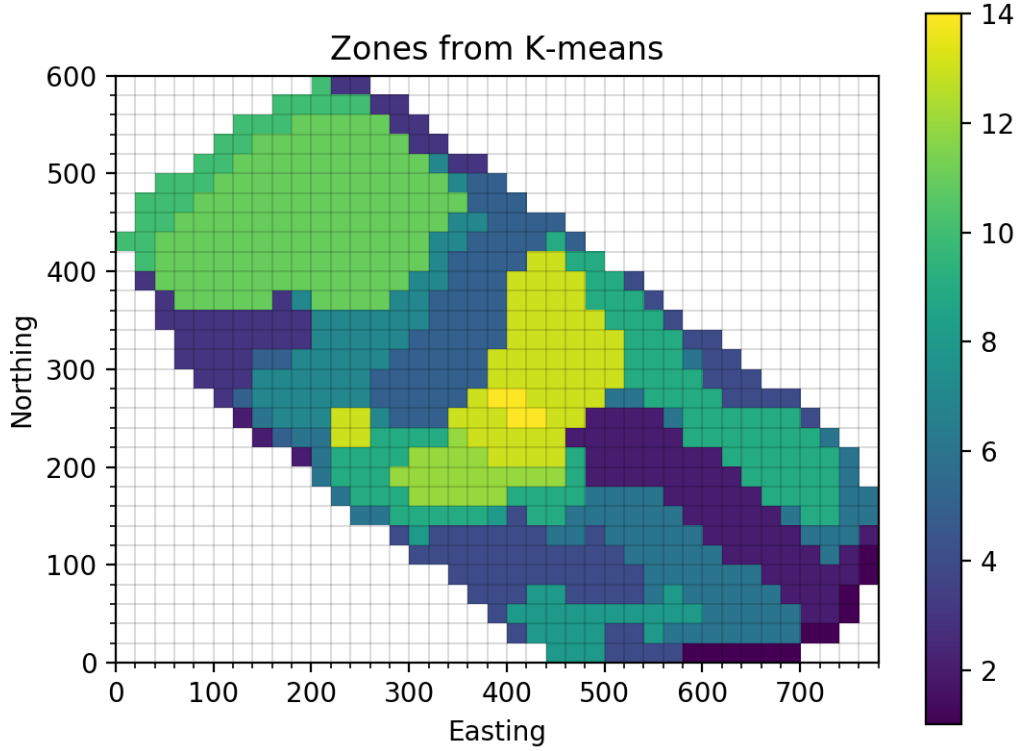



**Figure B2** k-means data clustering for LD field (k=15)

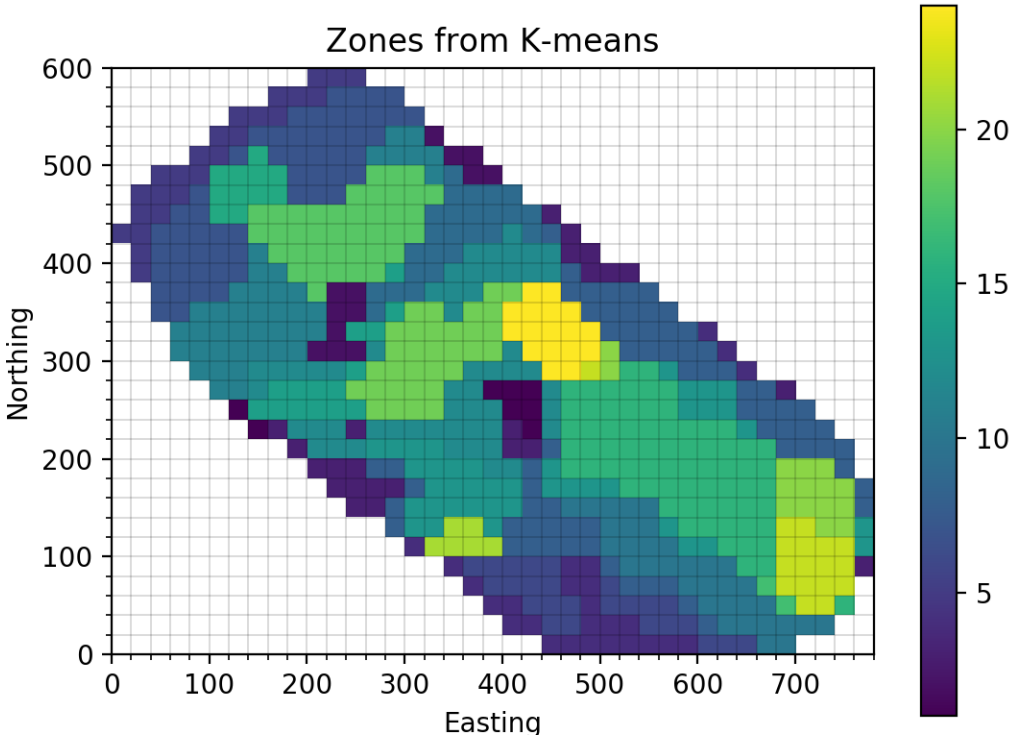



**Figure B3** k-means data clustering for LD field (k=25)

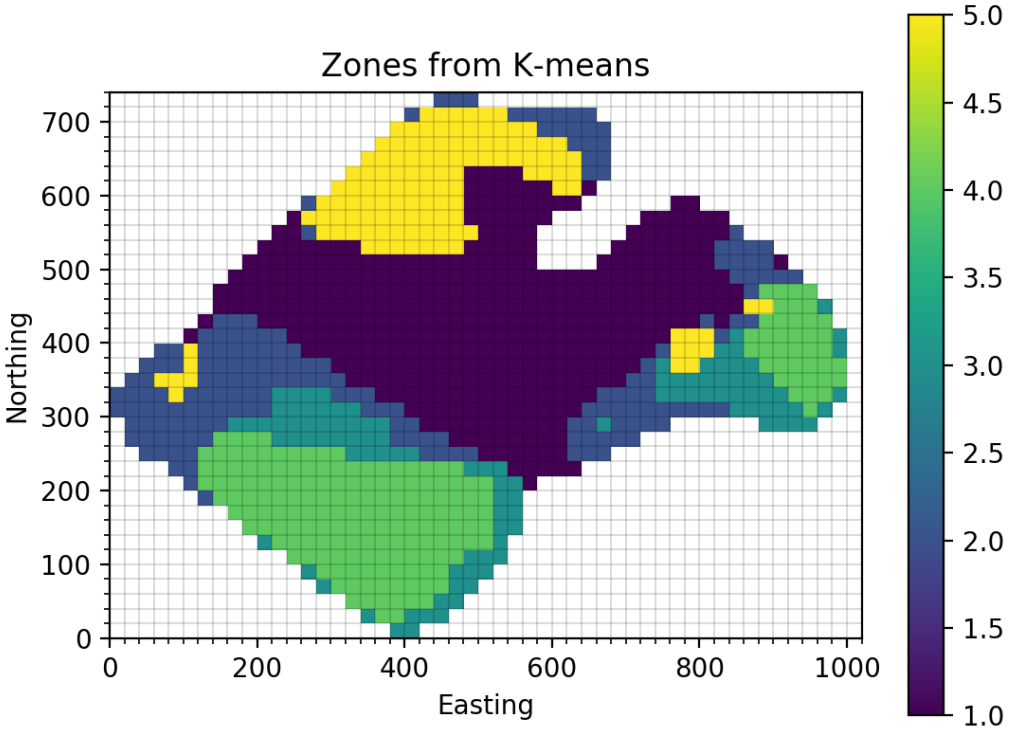



**Figure B4** k-means data clustering for RB field (k=5)

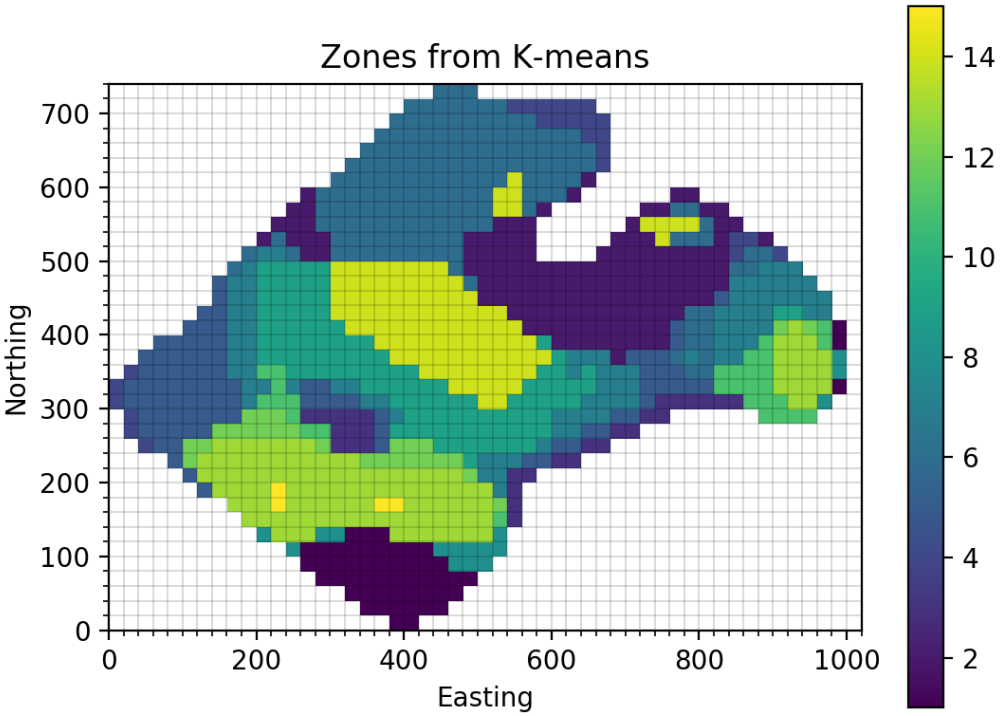



**Figure B5** k-means data clustering for RB field (k=15)

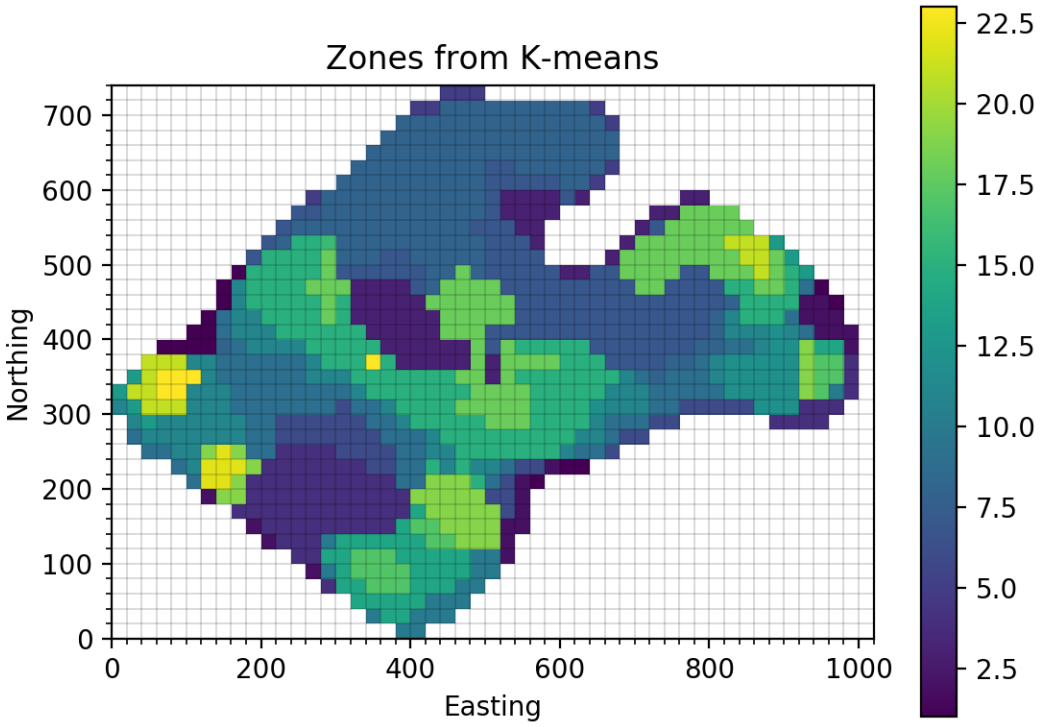



**Figure B6** k-means data clustering for RB field (k=25)

## C. Random forest modeling for Chapter 5

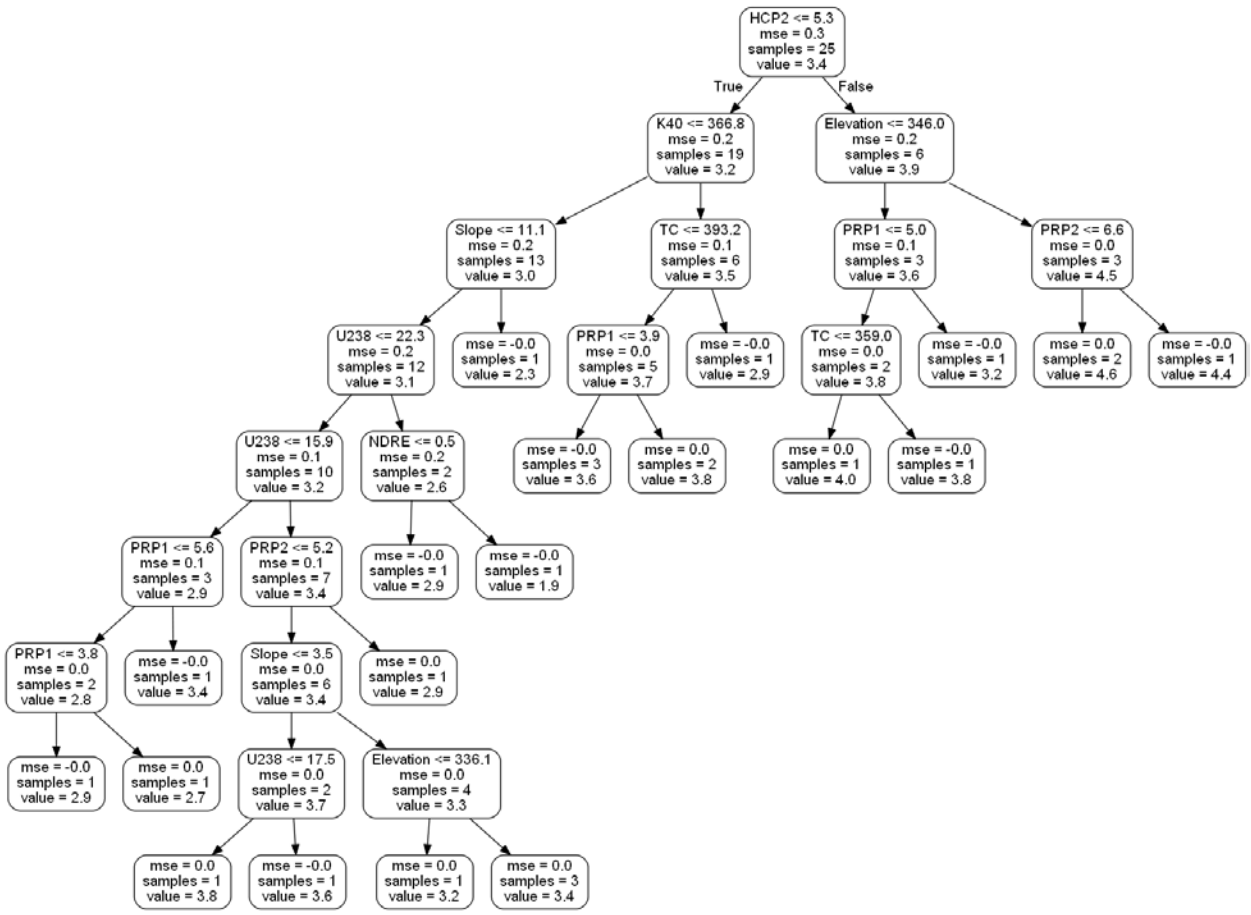



**Figure C1** Training (dataset split) and minimization of the node variance in the random forest model for soil SOM prediction.

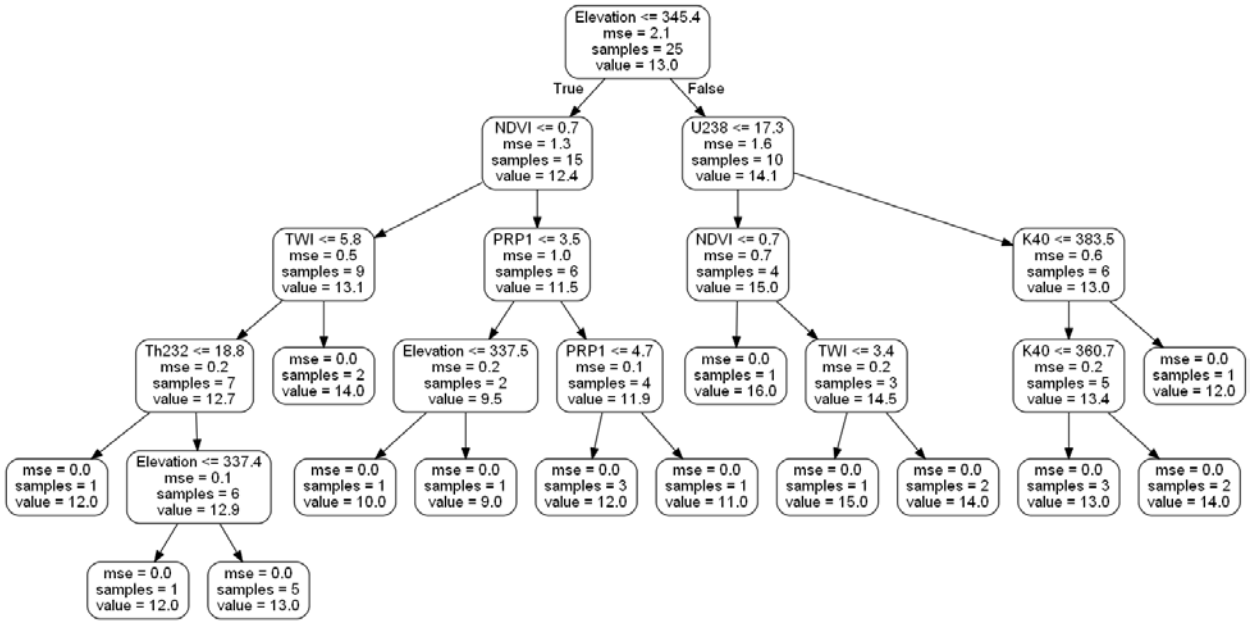



**Figure C2** Training (dataset split) and minimization of the node variance in the random forest model for soil CEC prediction.

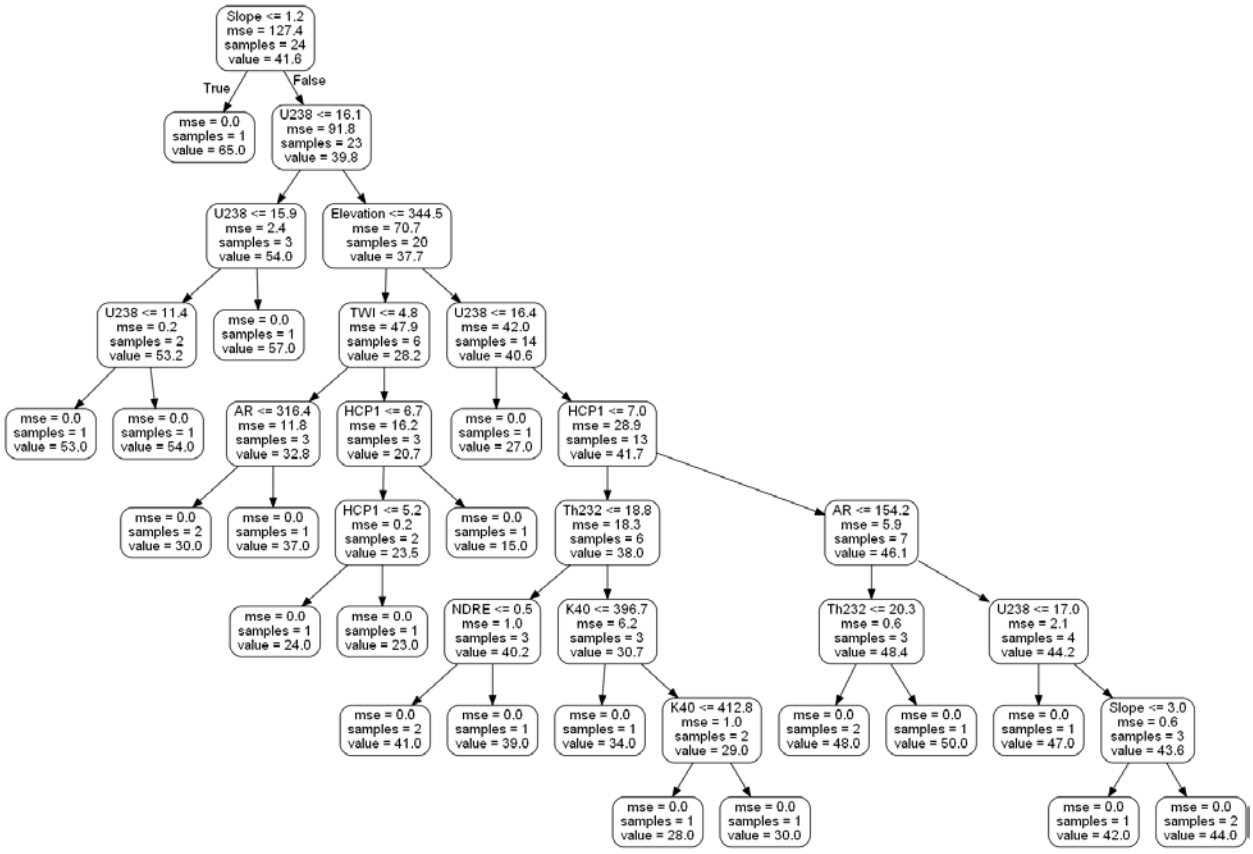



**Figure C3** Training (dataset split) and minimization of the node variance in the random forest model for soil P prediction.

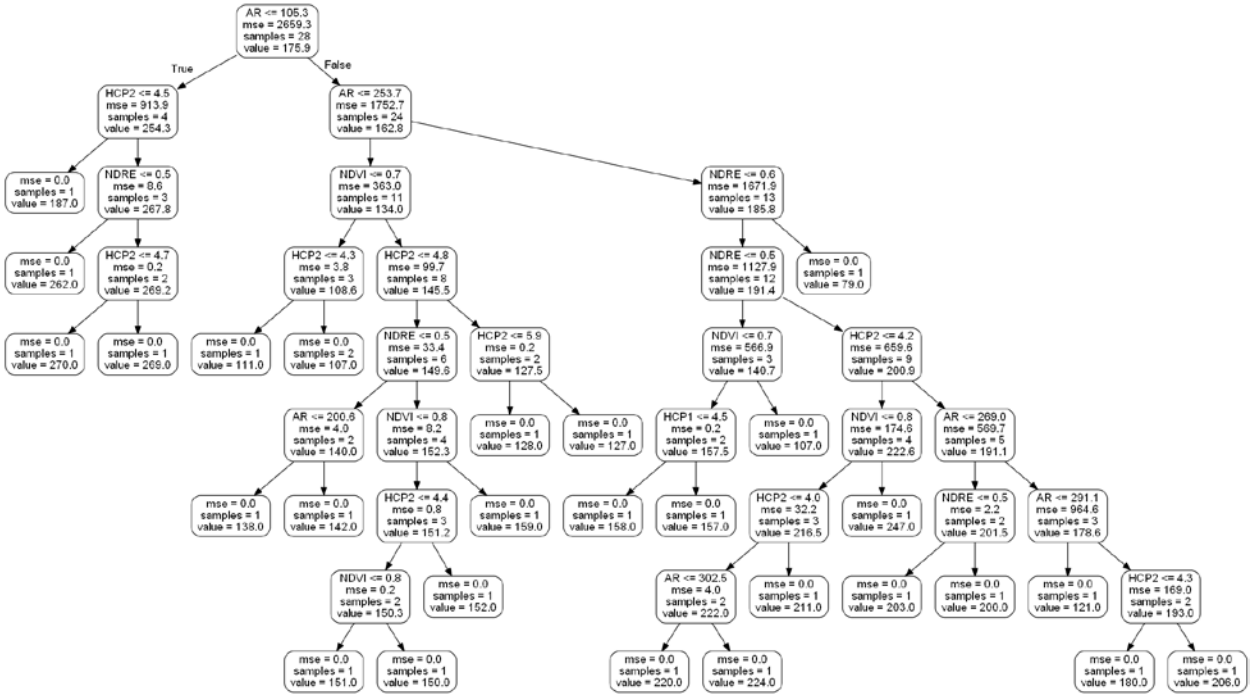



**Figure C4** Training (dataset split) and minimization of the node variance in the random forest model for soil K prediction.

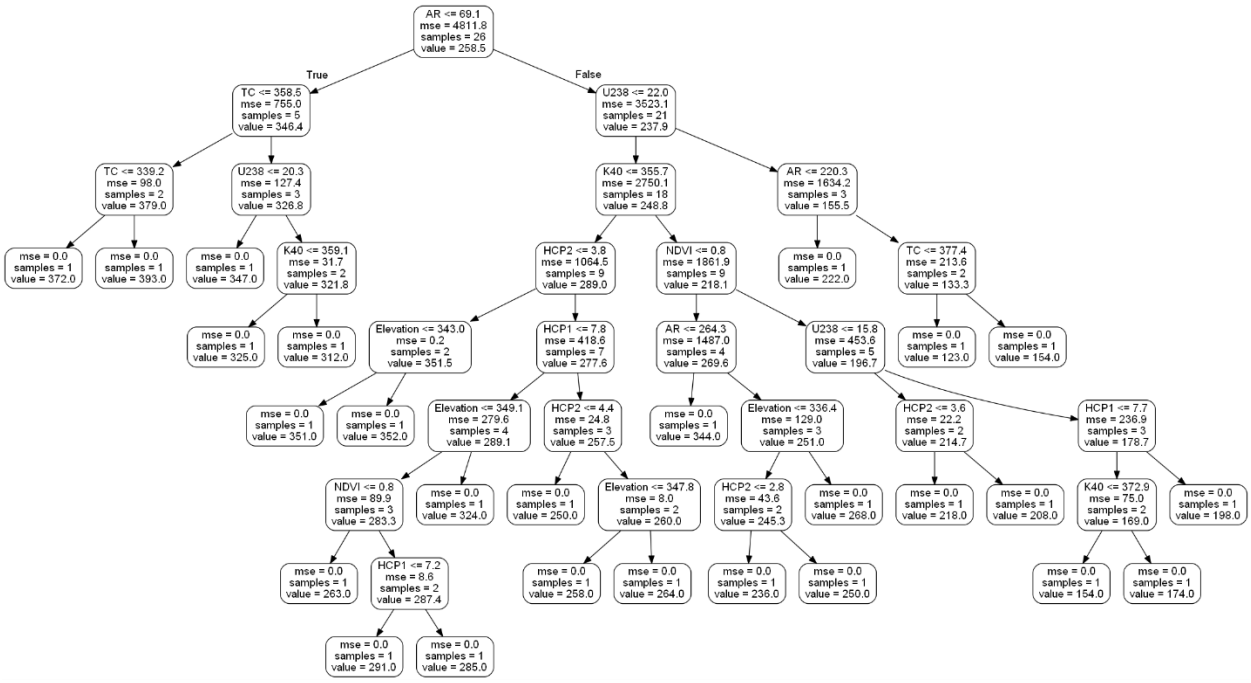



**Figure C5** Training (dataset split) and minimization of the node variance in the random forest model for soil Mg prediction.

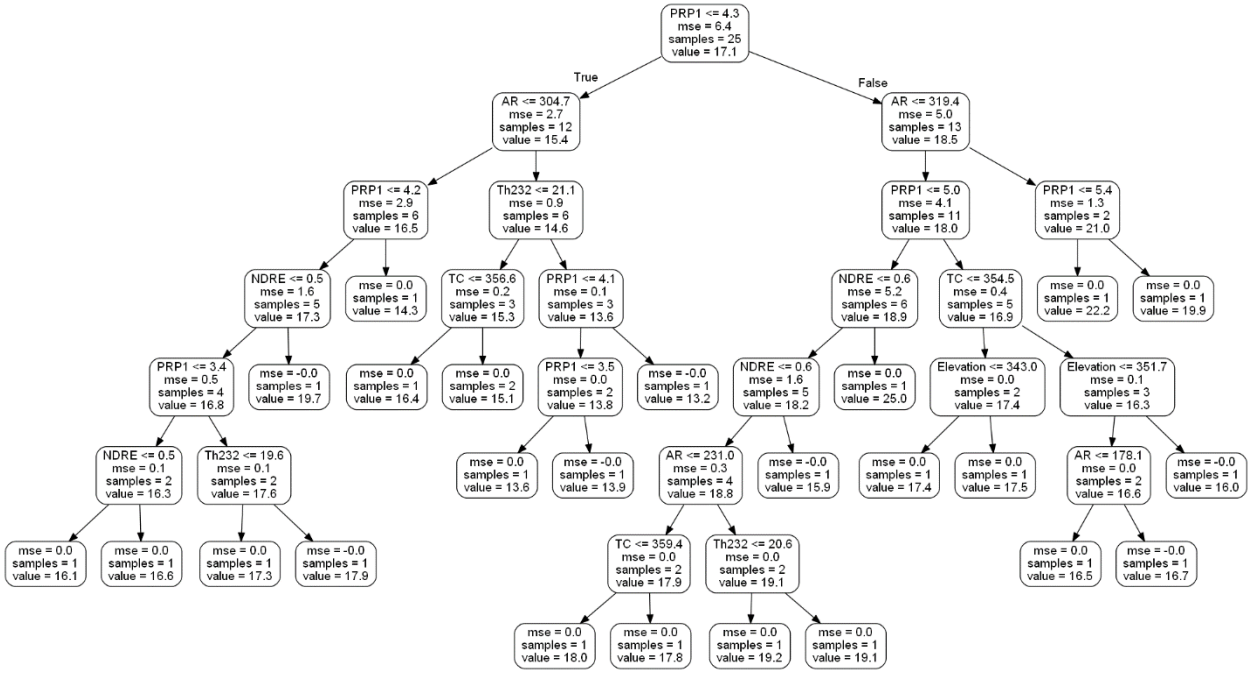



**Figure C6** Training (dataset split) and minimization of the node variance in the random forest model for soil Mn prediction.

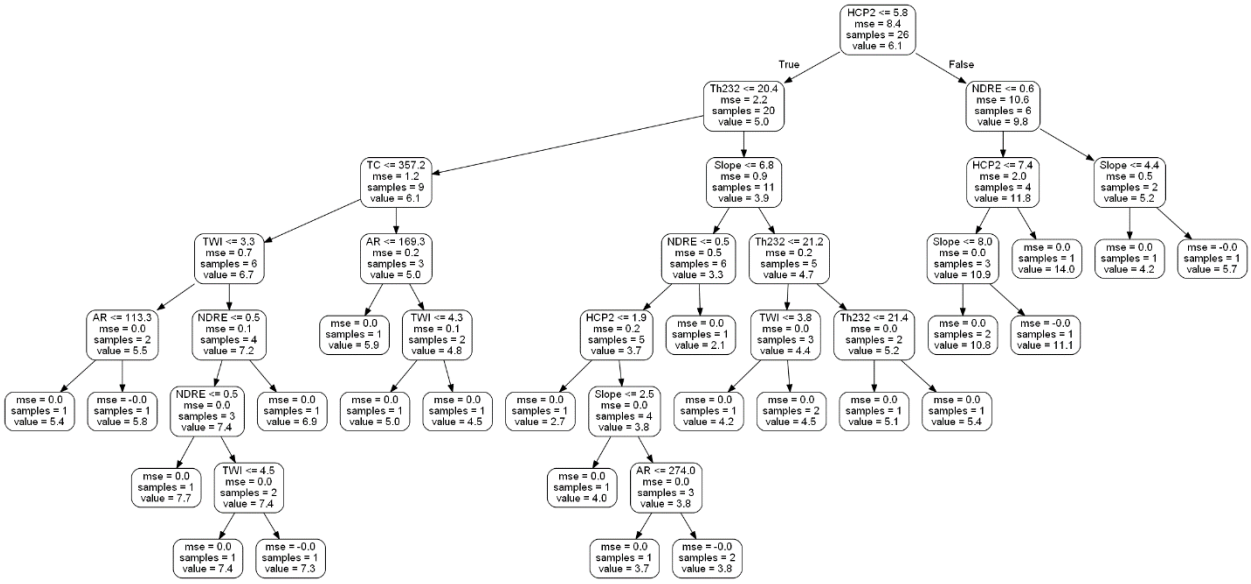



**Figure C7** Training (dataset split) and minimization of the node variance in the random forest model for soil Zn prediction.

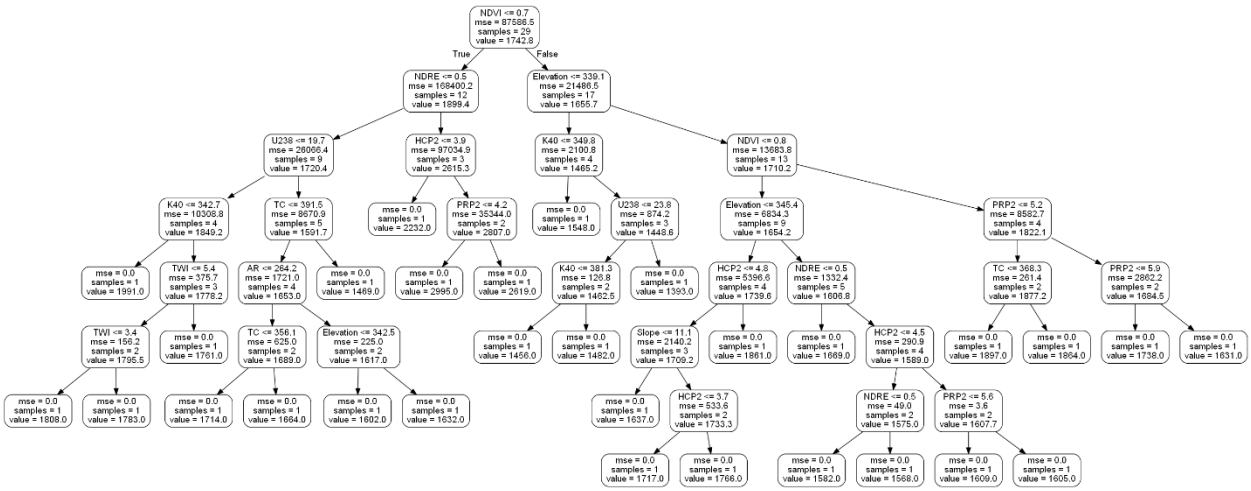



**Figure C8** Training (dataset split) and minimization of the node variance in the random forest model for soil Ca prediction.