# STATISTICAL AND MACHINE LEARNING METHODS FOR CROP YIELD PREDICTION IN THE CONTEXT OF PRECISION AGRICULTURE

By

Hannah Burdett

B.Sc. Hons., University of Windsor 2018

A thesis presented to Ryerson University

in partial fulfilment of the requirements

for the degree of

Master of Spatial Analysis

in the program of

Spatial Analysis

Toronto, Ontario, Canada, 2020

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Statistical and Machine Learning Methods for Crop Yield Prediction in the Context of Precision Agriculture, M.S.A., 2020, Hannah Burdett, Spatial Analysis, Ryerson University


## Abstract

It is of critical importance to understand the relationships between crop yield, soil properties, and topographic characteristics for agricultural management. This study's objective was to compare techniques to quantify the relationship between soil and topographic characteristics for predicting crop yield using high-resolution data and novel analytical techniques. The study was carried out across seventeen fields managed by a single cash cropping operation in Southwestern Ontario. Multiple linear regression, artificial neural networks, decision trees, and random forests were investigated to identify methods able to relate soil properties and crop yields on a point-by-point basis. Random forests were the most successful at predicting yield with an R-squared value of 0.93. Multiple linear regression was the least successful with an R-squared of 0.46. Machine learning techniques are often limited by their ability to extract meaningful relationships between variables. Thus, cross-validation techniques were applied to test the models and identify significant soil and topographic attributes when predicting yield.

## Acknowledgements

I want to express my sincere gratitude to my supervisor, Dr. Christopher Wellen, for providing this thesis with your invaluable expertise and time. He consistently allowed this paper to be my own work but steered me in the right direction whenever he thought I needed it. Thank you to Melissa Luymes and the farmers for compiling such high-resolution data to make such research possible. I want to thank Dr. David Atkinson and Dr. Wayne Forsythe for taking the time to be co-readers of my thesis. I would also like to thank Dr. Shuguang Wang for being the chair of my defense committee. Last but not least, I would like to thank my family, friends, and boyfriend for supporting me with my academic studies.

# Table of Contents

# List of Tables

# List of Figures

# List of Acronyms

ANN – Artificial Neural Network

ATP - Adenosine triphosphate

CEC – Cation Exchange Capacity

CPANN – Counter-propagation Artificial Neural Networks

DEM – Digital Elevation Model

DL – Deep Learning

DT – Decision Tree

DTPA – Diethylenetriaminepentaacetic acid

GIS – Geographical Information Systems

GPS – Global Positioning System

K – Potassium

ML – Machine Learning

MLR – Multiple Linear Regression

N – Nitrogen

OM – Organic Matter

OMAFRA – Ontario Ministry of Agriculture, Farming, and Rural Affairs

P – Phosphorus

PA – Precision Agriculture

RF – Random Forest

SKN – Supervised Kohonen Networks

SVR – Support Vector Regression

SWOOP – Southwestern Ontario Orthophotography Project

Zn – Zinc

# 1. Introduction

Agronomic scientists have conducted extensive research to map, monitor, analyze, and manage yield variability to optimize crop yield (Miao et al., 2006). One technique that assists with crop management is the use of crop yield predictions. Crop yield predictions are applied by crop managers to reduce losses by recognizing areas or factors that may cause adverse growing conditions, such as crop responses under climatic stress or deficiencies in nutrients. Additionally, crop yield predictions could be used to evaluate the optimum growing conditions so that fields may achieve their full growth potential (Dahikar and Rode, 2014). Although beneficial, crop yield depends on many interrelated factors such as e.g., elevation, cation exchange capacity, and soil nutrients (Miao et al., 2006). Predicting crop yield can be difficult as there is an increasing recognition that relationships between ecological drivers and their responses are commonly complex due to the non-linear interrelated factors (D'Amario et al., 2019; Gonzalez-Sanchez et al., 2014).

Several techniques have been used to understand the relationships between crop yield and soil or landscape properties (Miao et al., 2006). Statistical models such as multiple linear regression (MLR) have been widely considered (Drummond et al., 1995; Drummond et al., 2003; Khakural et al., 1999; Kravchenko & Bullock, 2000). However, results from MLR are often not satisfactory as MLR is limited to describing linear relationships between crop parameters and site variables, and the results are potentially misleading when these relationships are not linear (Drummond et al., 2003; Liu et al., 2001). For instance, Sudduth et al. (1996) used linear techniques on a dataset consisting of several site-years of topographic, soil, and yield data. They found that linear methods

generally failed to produce reasonable approximations of spatial yield variability, even with sub-field regions thought to be reasonably homogenous (Drummond et al., 2013). Multiple linear regression techniques using polynomial and interaction terms have also been considered (Kitchen et al., 1999) with some improvement over strictly linear models (Drummond et al., 2013).

Machine learning (ML) techniques have been applied through various agricultural systems over the past decade to provide more accurate solutions, primarily because of its ability to handle highly complex non-linear agricultural problems (Tantalaki et al., 2019). Unlike traditional statistical methods, ML does not make assumptions about the correct structure of the data model, such as the functional form, probability distribution, or smoothness (Khazaei et al., 2008; Mittal and Zhang, 2000; Seyhan et al., 2005). Instead, ML techniques have the ability to learn the relationship between dependent and independent variables through the data (Mittal and Zhang, 2000). ML techniques are based on semiparametric and nonparametric structures, with validation relying on prediction accuracy (Gonzalez-Sanchez et al., 2014). A decision tree is a nonparametric approach for building classification models. A nonparametric approach does not require any prior assumptions about the form of probability distributions that the class and other attributes satisfy. Applications of such techniques have been producing higher accuracy models from complex natural systems with multiple inputs (Dahikar and Rode, 2014).

Some comparisons among linear and ML techniques for crop yield prediction have been made. Gonzalez-Sanchez et al. (2014) compared multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression, and k-nearest neighbour for fields in Mexico. The dataset included records

from the fall-winter season in the years 1998-2006 and included a total of 6217 observations. However, this study focused primarily on the effect of climatic data on predicting crop yield and excluded several soil properties and topological characteristics. Primarily, cation exchange capacity (CEC) and elevation were not included as potential predictors. However, they were identified as among the top four most important soil and landscape factors for both corn yield and quality at two fields near Paris in eastern Illinois, USA, covering a total area of 40 ha by Miao et al. (2006). It has been established that soil properties, topographic characteristics, and crop yield exhibit spatial variability in agricultural fields (Corwin and Lesch, 2003; Jung et al., 2006; Metwally et al., 2019). For instance, Kravchenko and Bullock (2000) conducted a study on eight fields with soil data from 1994 to 1997 in central Illinois and eastern Indiana to evaluate how useful topographical information and soil properties can explain yield variability. The cumulative effect of the topographical features explained about 20% of the yield variability, while soil properties explained approximately 30% of the yield variability. Thus, spatial variability of soil and topographic properties can account for approximately 50% of agricultural production (Dahikar and Rode, 2014).

Precision agriculture (PA) utilizes variability in soil and topographic properties to manage fields through site-specific management strategies. Fields with a higher degree of spatial variability are likely to benefit from such crop management strategies (Mzuku et al., 2005). PA uses a wide range of technologies to gather, process, and examine data to guide targeted actions that advance the efficiency, sustainability, and productivity of agricultural practices (Tey & Brindal, 2012). This management strategy has the potential

to drive a new wave of increased agricultural productivity as well as contribute to the environmental sustainability of farming practices (Robertson et al., 2007).

Furthermore, previous research comparing linear and ML techniques often only provide insight into the efficacy of the model's ability to predict yield (Drummond et al., 2003). Although identifying the most useful predictive technique is beneficial for PA practices, comparing such methods does not inherently improve farm management practices. There is a lack of understanding of which factors influence yield as well as which attributes for yield predictions are most important. For instance, Drummond et al. (2003) compared artificial neural networks, stepwise multiple linear regression, and projection pursuit regression. Through this comparison, Drummond et al. (2003) identified artificial neural networks as the most effective method for crop yield prediction. However, more work is needed to link ML techniques to better decisions, not only to showcase their predictive abilities. For example, while techniques are compared within this study, additional cross-validation techniques have been applied to gain insight into the effect of soil and topographic attributes on yield. Through these cross-validation techniques and yield maps, the ML models assist in producing crop management recommendations.

The objectives of this study were to 1) evaluate the predictive ability of MLR, artificial neural networks (ANN), random forests (RF), and decision trees (DT) to identify which techniques provide the most accurate predictions, 2) identify which fields in operation have similar relationships and which fields differ, and 3) identify which variables may be the most important for yield. The study was conducted on a multiple site dataset, which included 145,500 observations of corn and soybean yield, topographic,

and soil nutrient characteristics. The attributes considered for this study included pH, soil organic matter (OM) content, cation exchange capacity (CEC), phosphorus (P), zinc (Zn), potassium (K), elevation, and topographic wetness index. Aside from identifying the most effective method for crop yield predictions, a set of cross-validation techniques were utilized to access the predictive ability for each field as well as groups of missing fields. Variables relationship with yield provided insight into limiting factors for crop growth within the fields. Additionally, variables of most importance were compared to yield maps to gain insight into spatial variability to assist in potentially guiding farm management practices.

# 2. Literature Review

## 2.1 Precision Agriculture

PA (or site-specific agriculture) utilizes rapidly evolving electronic information technologies to modify land management in a site-specific manner as conditions change spatially and temporally (Corwin and Lesch, 2003). PA aims to improve crop production while reducing adverse environmental effects. First developed in the mid-1980s, the technological pieces needed to bring PA came to be in the mid-1990s with the maturation of global positioning systems (GPS) and geographic information systems (GIS) (Corwin and Lesch, 2003). As such, PA is a technologically driven system that utilizes various technologies to assess soil, site, and crop variability (Pathak et al., 2019; Pierce and Nowak, 1999). Such technologies complement the observations of farmers and add an extra dimension to assessing agricultural systems' performance. Rigorous modern data mining techniques can find relationships and associations between multi-source observations that farmers can use to improve their farming practices (Cock et al., 2011; Delmotte et al., 2011; Lacey, 2011). For instance, this data-driven strategy uses site-specific knowledge to estimate proper fertilizer and pesticide application. This management strategy has the potential to drive a new wave of increased agricultural productivity as well as contribute to the environmental sustainability of farming practices. In some farming communities, several PA innovations have been introduced as standard practices. PA practitioners often combine soil nutrient data with the spatial variability of crop plants to prompt a targeted response to unfavorable crop or field conditions (Cook and Bramley, 1998; Robertson et al., 2007).

A review conducted by the US PA Dealership Survey in 2017 revealed that 55% of respondents used manual GPS guidance, and 78% surveyed used GPS guidance with auto tractor control (Erickson et al., 2017). However, a 2014 Grains Research and Development Corporation survey in the US showed that the national average adoption of variable rate technology and yield mapping were 9.0% and 29% of the cropped area (Lowenberg-DeBoer and Erickson, 2019). Variable-rate applicators in PA is an area of technology that focuses on the automated application of materials to a given landscape. With variable-rate technology, producers can maximize growth opportunities by tailoring seeding and crop nutrition applications to specific parts of their fields (Robertson et al., 2012). Hence, variable-rate technologies can reduce the amount of fertilizer and pesticide required to achieve a given yield, potentially benefitting farm profitability and reducing the environmental impact of crop production (Cook and Bramley, 1998; Dobermann et al., 2004; Jochinke et al., 2007; Rainbow and Wells, 2004; Schieffer and Dillion, 2013). Technologies used to implement variable-rate applications include GPS positioning, yield monitors, and variable-rate applicators (Cook and Bramely, 2001). GPS-based guidance technology can be used for many field operations such as sowing, tilling, planting, cultivating, harvesting, and weeding (D'Antoni et al., 2012). A common GPS navigation technology is auto-steer, which is an automated steering and positioning system (D'Antoni et al., 2012). Auto-steer could improve operator performance by reducing fatigue and increase the efficiency of the farm input application (Castle et al., 2015). The auto-steer system prevents human error, such as skipping and overlapping, which can lead to misapplication of pesticides, fertilizers, and seed. Auto-steer technology may also aid in reducing fuel consumption and emissions (Mishra et al., 2005; Chang et al., 2011).

PA technologies provide a foundation to collect big data. Big data is an evolving term that refers to voluminous structured, semi-structured, and unstructured data (Yadav et al., 2015). Structured data adhere to a predefined model of data as it typically follows a tabular format with relations between the different rows and columns. In contrast, unstructured data does not have a predefined data model, which may lead to inconsistencies and ambiguities that can be difficult to understand across conventional programs. PA utilizes a range of techniques and technologies that require new forms of integration to uncover hidden values from complex and diverse datasets (Yadav et al., 2015).

Furthermore, Stubbs (2016) suggests that the big data about agriculture are less about the size of the data than the combination of technology and advanced analytics, which creates a new way of processing information. Coble et al. (2016) supported this concept by defining the data in terms of volume, velocity, variety, and veracity. Volume refers to the quantity of data collected, Shearer (2014) puts the volume of PA data into perspective by reporting the data collected from planters via telematics in which 5.5 megabytes of data on location, cultivar, speed, and other geospatial and meta-data are collected for each acre. During planting seasons, the size of the aggregated PA data continues to accumulate. While agricultural operations are seasonal, crops such as corn, soybeans, cotton, rice, and wheat have varying peak planting times throughout the year. Thus, data velocity, the speed at which the data are produced, are collected over several months of the year rather than all at once (Yadav et al., 2015).

Moreover, many field operations such as tillage, spray applications, and harvesting occur at various times through the growing season, and each operation

contributes to the span of data collection (Shearer, 2014). The variety reflects the spectrum of data sources and inconsistencies within data structure and design (Coble et al., 2016; Yadav et al., 2015). Data may be collected using manual methods such as soil sampling, near-automated data collection for PA and the transfer from machine-based sensors and telematics. Other PA data may be collected and stored in a wide variety of unstructured formats, such as dates or descriptions. Veracity refers to the biases, noise, and abnormality of the data (Coble et al., 2016; Yadav et al., 2015). Data quality has been a contentious topic in PA, particularly regarding raw yield monitoring and data collected by PA technology sensors. For instance, the calibration of a yield sensor or combine operator may impact the veracity of yield data (Shearer, 2014).

ML applications are a key benefit of the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for big data analytics (Najafabadi et al., 2015). There is extensive literature on new techniques, such as ML, that have been used to manage big data in various PA applications (Ali et al., 2015; Chlingaryan et al., 2018; Liakos et al., 2018; Verrelst et al., 2015). For instance, Ruß (2009) evaluated various regression techniques to find suitable models achieving high precision and generality in terms of predictive yield capabilities. Neural networks, despite their site-dependency, proved robust, the support vector regression (SVR) used in the study was computationally less demanding and more accurate. In Were et al. (2015), SVR and artificial neural networks models were used to map the trends of soil organic carbon stocks, and the authors argued for the importance of data quality. Although ML was previously implemented in prior PA research, earlier studies, such as Ruß (2009) and Were et al. (2015), involved the incorporation of climate variables and the

implementation of remote sensing techniques for collecting yield and soil nutrient data. Few PA studies have used within field measurements collected through turbines or sampling. Rather the majority of PA studies are dependent on vegetation indices for soil nutrient and yield values (Drummond et al., 2013).

As previously mentioned, soil and topographic characteristics can account for 40% of agricultural production (Dahikar and Rode, 2014). A further understanding of these attributes' spatial variability and their predictive yield capabilities can help establish comprehensive farm management plans for PA practices. A thorough understanding of spatial variability can, for example, aid in the composition and application of fertilizers to aid with the optimal growing conditions. Additionally, there is a lack of research utilizing sensor yield values and soil samples when estimating yield in ML studies.

### 2.2 Soil Management

The soil is a heterogeneous matrix with a wide variety of physical, chemical, and biological characteristics (Soil & Test, 2017). Agriculture is one of the most impactful anthropogenic practices affecting soil properties and, consequently, their functions. Properties of soil allow researchers to gain insight into ecosystems' dynamics and the impingement in agriculture (Liakos et al., 2018). Knowledge of the spatial variability of soil helps to understand crop production variability (Tantalaki et al., 2019). For instance, nutrient interaction may have a positive or negative influence on crop yield. If the combination of nutrients results in a growth response more significant than the sum of individual effects, the interaction is positive (Fageria, 2001). Precise soil property estimates are necessary for optimal soil management, nutrient planning, and land-use decisions (Lahoche et al., 2003). For instance, soil moisture plays a critical role in crop

yield variability. Monitoring soil moisture enhances the understanding of the water exchange rate at the atmosphere/ground interface (Pasolli et al., 2011). Yield variability is one of the most important within field characteristics that influences the adoption of PA technologies (Paxton et al., 2010; Zhang et al., 2002). The crop yield is mostly influenced by the availability and allocation of soil nutrients in a field (Zhang et al., 2002). Thus, effective management of soil nutrients depends on the producer's ability to capture the distribution of soil nutrients (Asare and Segarra, 2018). Soil sampling techniques, such as grid sampling, help monitor soil nutrient variation in-field and store it as a soil test map (Wollenhaupt and Wolkowski, 1994). The information stored in soil test maps assisted in developing PA input management practices, such as fertilizer maps, on which variable rate application of agricultural inputs are based (Asare and Segarra, 2018; Erickson and Widmar, 2015; Fleming and Westfall, 2000; Franzen and Peck, 1995). A standard soil test would consist of pH, organic matter (OM), phosphorus (P), potassium (K), zinc (Zn), and cation exchange capacity (CEC).

Soil pH is one of the main factors, primarily affected by chemical crop inputs that influence nutritional availability, crop growth, and microbial diversity. A soil with a pH of 7.0 or higher is considered alkaline or basic soil. If the pH is less than 7.0, the soil is called acidic (A&L, 2011). As soils become increasingly acidic essential nutrients, such as P, become less available to plants. Low pH increases the availability of other elements, such as aluminum, which may lead to toxic environments for plants if the concentration of such nutrients increases. Soil acidity influences the availability of elements and directly influences the microbial population of the soil (South and Davey, 1983). Achieving optimum pH not only increases the availability of essential nutrients, but also

supplies additional calcium and magnesium, improves soil conditions for microorganisms, increases the effectiveness of triazine herbicides, and improves soil structure. However, human activity can impact the pH of soils; the addition of most nitrogen fertilizers and organic nutrient sources, such as compost and manure, leads to the formation of nitric acid and/or sulfuric acid. Both are strong acids that cause a decrease in the soil's pH (Bergstrom, 1987; Ristow, 2010). The soil pH ranges recommended for corn is 5.8 to 6.2, while soybeans have a recommended pH range of 6.6 to 7.0 (Ristow, 2010).

Soil OM is the proportion of soil composed of plant or animal tissue at various decomposition rates and a biological measure of healthy soil. OM is composed primarily of plant residues and live microbial biomass, detritus, and humus. The living microbial biomass contains the microorganisms responsible for the decomposition of both plant residues and active organic or detritus soil. Humus is the stable fraction of organic soil matter produced from the decomposed tissue of plants and animals. Both types of OM contribute to soil fertility as the breakdown of these fractions results in the release of plant nutrients such as nitrogen, phosphorus, potassium, etc. The fraction of the humus has less impact on soil fertility since it is the final decomposition product (Ketterings et al., 2003). However, soil fertility management is still vital as it contributes to soil structure, soil infiltration, and CEC. In particular, OM increases the CEC of soil or its capacity to maintain and supply vital nutrients such as calcium, magnesium, and potassium over time. Additionally, OM enhances soil ability to withstand changes in pH, often referred to as buffering energy. OM intensified the overtime decomposition of soil

minerals, making available the nutrients in the minerals for plant uptake (Bergstrom, 1987; Ketterings et al., 2003).

Phosphorus, next to nitrogen (N), is a principal yield-limiting factor for annual crop and forage production in acid and alkaline soils of temperate regions (Fageria, 2006; Hodgson et al., 1966; Robson and Pitman, 1983). Hence, evaluating P interaction with other nutrients is essential to maintain a sufficient supply of nutrients to increase crop yields (Fageria, 2006). The primary function of P in a plant is storing and transferring photosynthesis-generated energy for use in growth and reproductive processes. It is an essential component of adenosine triphosphate (ATP). ATP is involved in most biochemical processes in plants and allows nutrients to be extracted from the soil (Ketterings et al., 2003). Additionally, it is also important in the growth of cells and the production of DNA. Adequate levels of P facilitate root growth and winter hardiness, promoting tillering, and accelerate maturity (Elrashidi, 2010). It generally has a significant positive relationship with the absorption of N and the growth of plants. It is often accepted that increased growth required more of both N and P, with the assumption that mutually synergistic effects result in the simulation of growth and enhanced absorption of both elements (Sumner and Farina, 1986). There are several different forms in which it can exist in the soil; plants available inorganic, organic, absorbed, and primary mineral P (Ketterings, 2003). Additional crop management strategies are applied to optimize crop uptake of available P as climatic and site conditions can influence its mineralization rate from OM decomposition. Cool climates often result in slower OM decomposition, releasing P more slowly. Furthermore, it is released faster in well-aerated soils and slower in well-saturated wet soils (Elrashidi, 2010). pH inherently influences P-

availability with pH values between 6 and 7.5 considered ideal for P-availability. While pH values below 5.5 and between 7.5 and 8.5 limits P-availability due to aluminum, iron, and calcium (Elrashidi, 2010). For instance, plant-available inorganic P reacts with dissolved iron, aluminum, or manganese in acidic soils or calcium in alkaline soils to form phosphate minerals (Ketterings, 2003). It does not readily each out of the root zone; instead, a potential loss of P is mainly associated with erosion and runoff. The soil solution has a very low concentration of soluble phosphate, and it is essentially immobile. Thus, crops' initial uptake of P comes from soil solutions (Elrashidi, 2010). Management strategies include applying fertilizer or manure to the soil to increase available P.

After N and P, K is the most likely the limiting nutrient for plant growth and is unique in the diversity of roles it plays in plant metabolism processes (Dibb and Thompson, 1985). Enzyme activation, acting as an osmoticum to maintain tissue turgor pressure, regulating the opening and closing of stomates, and balancing anion charge is the physiological function of K in plant cells (Pettigrew, 2008). Additionally, K fertility management is of importance as plants with optimum K levels are more resistant to environmental stresses, including drought (Ketterings et al., 2003). It is available through three major pools: soil mineral K, fixed K, exchangeable K, and soil solution. Soil mineral K is not available for plant uptake. However, as soil minerals break down over time, K is released to the soil solution. Fixed K is a component of the soil's internal clay mineral structure. This pool is accessible gradually over time for plant uptake; however, the total amount of K in solution is relatively small. Much of the K required for crop production comes from K in soil solution and exchangeable K. Due to the positive charge

of the K ion, it is attracted to soils negative charge and does not readily leach (Ketterings et al., 2003).

Zinc is a micronutrient that plays a critical role in several key plant physiological pathways to function normally (Alloway, 2002; Mousavi et al., 2011; Yosefi et al., 2011). Zinc is a constituent in enzymes involved with photosynthesis and an important element in the carbohydrates metabolism (Yost et al., 2011). Soils deficient in Zn may result from high pH as Zn becomes less available as pH increases. A deficiency in Zn generally occurs at a pH greater than 7.4 but may occur at a pH as low as 6.5. Additional factors that influence Zn concentrations are soils with low OM and restricted root growth (Yost et al., 2011).

Cation exchange capacity measures the soil's ability to hold positively charged ions. It is a very important soil property influencing soil structure stability, nutrient availability, soil pH, and the soil's reaction to fertilizers (Hazleton and Murphy, 2007). The clay minerals and OM components of soil have negatively charged sites on their surfaces, which adsorb and hold positively charges ions, cations, by electrostatic force. This electrical charge is critical to the supply of nutrients to plants as many nutrients exist as cations. In general terms, soils with large quantities of negative charge are more fertile as they retain more cations (McKenzie et al., 2004). The main ions associated with CEC in soils are the exchangeable cations such as calcium, magnesium, sodium, and potassium (Rayment and Higginson, 1992). The CEC of soils varies according to the percentage of clay, the type of clay, soil pH, and quantity of OM (McKenzie et al., 2004). Soils with low CEC are more likely to develop deficiencies in potassium, magnesium, and other cations (Ketterings, 2007). The fertility of soils decreases with decreasing pH, and the

lower the CEC of the soil, the faster the soil pH will decrease over time. The addition of OM or other management processes such as liming will increase the CEC of the soil, although it may take several years to take effect (Moore et al., 2001).

### 2.3 Topographic Properties

Topography plays an important role in agricultural fields in shaping the spatial variability of soils, surface and subsurface hydrology, and crop yield (Iqbal et al., 2005). Landscape topography influences the erosion and/or deposition of soil particles and soil nutrients, with resulting changes in physical and chemical properties of uphill and downhill soils (Ovalles and Collines, 1986; Pennock and Jong, 1990). Kravchenko and Bullock (2000) reported topography explained approximately 30% of the observed variability in OM, P, and K concentrations. Furthermore, Gburek and Sharpley (1998) discussed how P's loss from land to stream is primarily regulated by the interaction of P sources, such as soil, crop, and land management, with its transport factors. Transport factors for P may include runoff, erosion, and channel processes.

Additionally, landscape topography affects water availability due to both horizontal and vertical water redistribution (Verity and Anderson, 1990). Water redistribution plays a critical role in the amount of water available to plants which has a significant impact on field yield variability (Afyuni et al., 1993; Daniels et al., 1987; Fiez et al., 1994; Holt et al., 1964; Hanna et al., 1982; Wright et al., 1990). Li and Lindstom (2001) reported water erosion as the primary cause for the overall decline in soil quality on a steep cultivated hillslope, while tillage erosion had a similar contribution to the overall level of soil quality on a terraced hillslope.

Topography-yield relationships have been studied extensively. Mahler et al. (1979) used 3-by-30-m plots located at ridgetop, bottomland, and south slope positions to examine the relationships between topography and dry pea yield. They found higher yields and higher soil water contents at the bottomland landscape position. Ciha (1984) evaluated wheat yields from individual parcels located at concave, middle, toe, convex, and interfluve sites, and found that landscape position was a significant factor affecting yield. Furthermore, aspect have significant correlations with wheat yields. Kravchenko et al. (2000) reported higher crop yield at lower slope locations, and a wide range of yield values on moderate and higher slopes during moderate to dry weather conditions. However, they found low yield values were measured on lower slope locations during the wet season. Additionally, Kravchenko et al. (2000) examined the effects of derived topographic and hydrologic indices on variability in soil properties and crop yield. They recorded a significant negative correlation between crop yield and elevation, slope, and curvature.

The development of GIS technology has made it possible to create digital elevation models (DEM) for terrain analysis (Da Silva and Silva, 2008). From these DEMs, several topographic attributes can be derived when evaluating yield variability. Wilson and Gallant (2000) divided topographic attributes into two categories: primary and secondary attributes. Primary attributes are computed directly from DEMs, whereas a combination of primary attributes determines secondary attributes. The most common primary attributes used in topography-yield studies are elevation, aspect, slope, upslope contributing area, and flow length (Da Silva and Silva, 2008). Elevation data are particularly useful for relating topography to soil properties (Moore et al., 1993; Odeh et

al., 1994). Secondary attributes are physically or empirically derived indices that describe different landscapes (Moore et al., 1991). These secondary attributes include flow accumulation, flow direction, wetness index, distance to flow accumulation lines, stream power index, and sediment transport index (Da Silva and Silva, 2008). Most studies only analyze the relationship between primary topographic attributes and yield variability (Bakhsh et al., 2000; Kravchenk et al., 2000; Kasper et al., 2003; Yang et al., 1998). However, the relationship between yield and secondary topographic attributes is found less frequently (Kravchenko and Bullock, 2000; Iqbal et al., 2005; Da Silva and Silva, 2006). Secondary attributes are significant as they can be used to quantify the role topography has played in the redistribution of water in landscapes. Thus, such knowledge can be used to study quantitative relationships between yield and topography, as well as yield and soil physical and chemical properties on large scales.

### 2.4 Crop Yield Predictions

Yield estimation is one of the most important issues in PA (Gonzalez-Sanchez et al., 2014; Pantazi et al., 2016; Ruß, 2009). Accurate and timely yield forecasting is necessary for decisions regarding storage, marketing, and transportation. As stated by Ruß (2009), yield prediction traditionally has relied on farmers long term experience for crops, specific fields, and climate conditions. Simple estimators, such as the average of several previous yields or the last obtained yield, are also used. Nevertheless, crop yields differ spatially and temporarily with a non-linear behavior that introduces large deviations from year to year and place to place within a field (Liu et al., 2001; Drummond et al., 2003; Schlenker & Roberts, 2006).

Two commonly used approaches for predicting crop yield responses include process-based modeling and statistical modeling. Process-based crop models are powerful tools for crop yield predictions as they simulate the physiological processes of crop growth and development in response to environmental conditions and management practices (Jeong et al., 2016). They provide computerized representations of crop growth, development, and yield, simulated through mathematical equations as functions of soil conditions, weather, and management practices (Basso et al., 2013; Hoogenboom et al., 2004). The strength of such models lies in its ability to extrapolate the temporal crop growth patterns and yield beyond a single experimental site. Process-based models are only an approximation of the real world, and many do not account for important factors such as diseases, weeds, insects, tillage, and phosphorus (Irmak et al., 2001). The models range from simple to complex. Simple models often utilized for yield estimation across large land areas based on statistical information related to climate and historical yields with little information about soil-plant systems. Whereas, more complex mechanistic models may provide detailed explanations of soil, plant, and atmospheric systems. Additionally, more complex models may require a large amount of input data, which may not be available (Basso et al., 2013). The calibration requirements of process-based crop models remain challenging for timely predictions of crop yield at a regional or global scale (Jeong et al., 2016).

Furthermore, statistical modeling estimates the direct relationship between predictor variables, such as soil factors and climate, and crop yield in a given data set without considering the underlying processes in crop physiology and ecology. Statistical models can provide simple but reasonable predictions, provided that sufficient and

reliable data has been used for training the model. Statistical models are more reliant on field calibration data and may provide commonly used performance assessment measures for uncertainty analyses. Simple and complex linear methods, including various forms of multiple linear regression, have been widely considered (Drummond et al., 1995; Kitchen et al., 1999; Khakural et al., 1999; Kitchen et al., 1999; Kravchenko and Bullock, 2000), with limited success. For instance, Landau et al. (2000) developed a regression model to determine the effects of climatic variables on wheat yield. Their final model had an r-value of 0.41 and provided insight into the most important explanatory variables and the weather effects they represent to be assessed. Sudduth et al. (1996) used linear techniques on a dataset consisting of several site-years of topographic, soil, and yield data. They found that linear methods were generally insufficient at producing good approximations of spatial yield variability, even within sub-field regions thought to be reasonably homogenous.

Urban et al., (2012) used statistical models to determine the effects of temperature increases on maize yield in the United States, concluding that temperature increases will play a meaningful role in yield decrease under climate change. In general, the findings of statistical models can not necessarily be extrapolated to other space and time due to differences in climates, soils, and weather not included in the population of information from which the statistical information was obtained. Furthermore, the applicability of this type of crop-weather model to areas outside the regression region is limited. A principal problem associated with statistical crop models is that yield simulations can be carried out outside the range of weather and technology information from which the model was developed. Statistical models can be used to inform other models and may provide

insights into past yields and historical influences (Basso et al., 2013; Gage and Safir, 2011; Lobell et al., 2011).

### 2.5 Machine Learning Models

In recent years different ML techniques have been implemented to achieve accurate yield predictions for different crops (Sudhadra et al., 2016). Predominantly the most successful ML techniques have been ANN (Drummond et al., 2003; Fortin et al., 2011; Liu et al., 2001; Safa et al., 2004), Support Vector Regression (Ruß, 2009), M5-Prime Regression Trees (Frausto-Solis et al., 2009; Marinković et al., 2009; Ruß and Kruse, 2010; Wang and Witten, 1997), and k-nearest neighbor (Zhang et al., 2010). One of the main advantages of ML techniques is that they are capable of autonomously solving large non-linear problems using datasets from multiple (potentially interconnected) sources (Chlingaryan et al., 2018). Inputs from different sensing systems, such as climatic characteristics or soil, can be combined to accurately predict yield and provide crop recommendations (Bendre et al., 2015). In real-world scenarios, ML enables better predictive actions without or with minimal human intervention. ML provides a powerful and flexible corporation of expert knowledge into the system. These are some of the key characteristics of the ML techniques that make them widely used in many domains and highly applicable to PA (Chlingaryan et al., 2018).

As soil and climatic conditions play an important role in crop growth and yield, Pantazi et al. (2016) predicted differences in wheat yield within the field using online multi-layer soil data and yield values computed from satellite imagery. Self-Organizing Maps and data from a single growing season were used in this study. They compared the performance of counter-propagation artificial neural networks (CPANN), Supervised

Kohonen Networks (SKN), and XY-fused Networks for predicting wheat yield. The SKN model had the best average overall accuracy of 81.65% compared to CPANN (78.3%) and XY-fused Networks (80.92%) models, showing that the SKN model had the best overall performance. In another study, Nari and Yang-Won (2016) applied four ML techniques, support vector machines, RF, extremely randomized trees, and deep learning (DL), to estimate corn yield in Iowa State. Comparisons of the validation statistics show that DL provided more stable results by overcoming the issue of overfitting.

Shearer et al. (1999) examined a considerable number of variables, including satellite imagery, soil conductivity, and fertility, for a relatively small number of site-year observations of data with limited success. ML models are more likely to be successful with large, diverse datasets. Thus, several years of data collection are often required in PA studies. For instance, Liu et al. (2001) used a standard backpropagation neural network to estimate corn yields over several years of a small data plot. Their findings were encouraging, with predictive error reports being approximately 20% of the actual yield; however, only a single validation set was used.

Despite significant recent developments in ML and successful implementation in several areas, ML techniques have some fundamental limitations when used naively in a purely data-driven fashion. The accuracy of the predictions and the uncertainties generated by the ML algorithms strongly depends on the quality of the data. The representativeness of the model and the dependencies between the input and target variables exist within the data. Data with a high level of noise, presence of outliers, erroneous data, biases in the data, and incomplete datasets will significantly decrease the model's predictive power. Several strategies, such as incorporating expert knowledge into

the covariance function, transfer learning, outlier detection, and model selection through automated cross-validation can be employed to overcome these limitations (Chlingaryan et al., 2018). Additionally, ML results are often considered opaque as they fail to be interpretable and explicable (Krishnan, 2019). While ML techniques often make very successful predictions, there is a lack of understanding of how these classifiers function and the relationship between the dependent and independent variables can be unclear.

# 3. Study Area

Data was collected on seventeen fields located in Southwestern Ontario. All the fields are owned by the same farmer and undergo similar farm management practices. The fields range from 5.17 to 80.04 acres in size. Cash crops were grown in a rotation of corn and soybeans, with no animal's present. The soil in this area is comprised of Brookston clay soil with subsoil claypan horizon(s) varying between silty clay loams, silty clay, clay loam, or clay (Richards, 1949). The topography of Southwestern Ontario is primarily flat, with slopes less than 1% (Frank and Ripley, 1990). Southwestern Ontario generally has a drier, warmer climate with sufficient soil fertility; thus, it is considered ideal for agricultural practices (Tan and Reynolds, 2003).

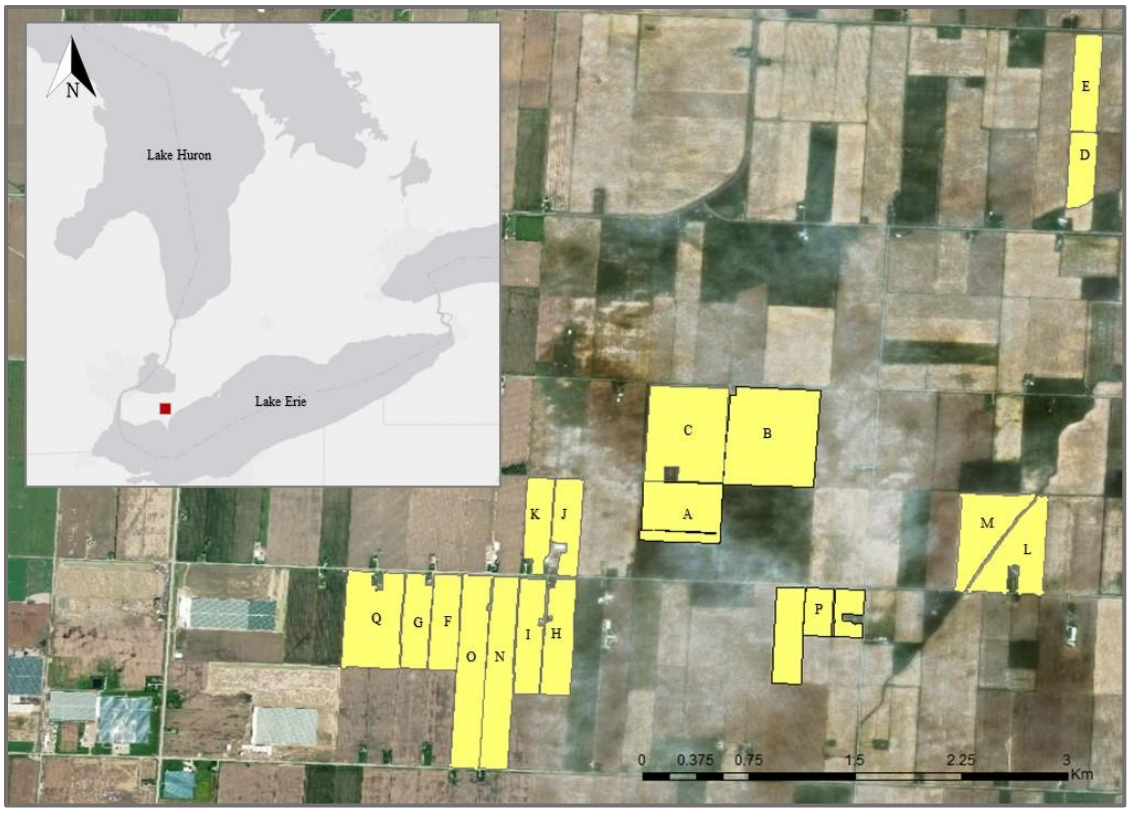



*Figure 1. The locations and shape of the seventeen fields located in Southwestern, Ontario.*

# 4. Data

## 4.1 Crop Yield and Soil Nutrients Data

The yield and crop data were purchased from the field's agronomists. In which, crop yield measurements for this study were obtained using a full-size combine equipped with a commercial yield sensing system and a global positioning system (GPS) receiver. For this study, yield data was collected for a single year in 2017. The yield sampling points were collected at an average distance of 2m apart with corn having significantly higher bushels per acre (bu/acre) yield values compared to soybeans. Table 1 shows the fields, the crop harvested, descriptive yield data, and the number of observations. The yield values were normalized using a Z-score to compensate for this difference in yield scales. The Z-score rescales the original variable to have a mean of zero and a standard deviation of one (Patro and Sahu, 2015).

Georeferenced grid soil samples were taken primarily in June and October of 2017, following the Southwestern Ontario sampling guidelines (OMAFRA, 2009). Grid soil sampling involved samples being collected in a systematic grid so that location information would be available for each sampling point. This sampling technique provided a spatial representation of soil nutrients throughout the fields. Soil samples were taken at an average distance of 110m apart, at a depth of 6-inches. A&L labs performed a chemical analysis of the samples using accredited techniques. Measured data (A&L, 2017) included pH, OM, P, K, Zn, and CEC. A&L is a soil fertility lab accredited by the Ontario Ministry of Agriculture, Food, and Rural Affairs (OMAFRA, 2009).

Soil pH is measured with a standard lab test using electrode and a saturation paste, in which the soil is crushed to make a saturated paste. pH electrodes are then inserted into the paste to determine the pH while slowly moving the electrodes within the paste. The soil pH was a measure of the activity of hydrogen ions in the soil solution. As previously mentioned, a pH of 6.9 or less is acidic, while soils with a pH of 7.0 are neutral; values higher than 7.0 are alkaline (A&L, 2017). OM was measured by directly measuring OM's weight loss from

the soil when it is burned, which is referred to as the loss on ignition. Samples are places in a muffle furnace overnight and the weights before and after ashing are compared. Additionally, soil colour was an indication of OM content as soil darker in colour has a higher portion of organic matter. The Olsen method, also referred to as the sodium bicarbonate method, was utilized to measure the amount of readily available P in alkaline soils. The extracting solution has a pH of 8.5 and so it is ideal for soils with a pH range from 6.0-8.0. The calcium phosphates in the soil and some of the organic phosphates dissolve in an extracting solution of weak sodium bicarbonate. Extractable K was determined using the Mehlich III method in which ammonium ions displace the K cations from the exchange sites. The concentration of the K cations was then measured in the extract. Zn was based on a diethylenetriaminepentaacetic acid (DTPA) extraction. For this extraction, the soil was mixed with a 0.005 M DTPA solution at a ratio of 1-part soil to 2-parts solution and shaken for an hour. The Zn in the soil is complexed by the DTPA and held in solution. Following extraction and filtering, the Zn content is the extract is measured. A soil's CEC depends upon the quantities and types of clay minerals and OM present. For instance, soils with high CEC will generally have higher levels of clay and OM. CEC was measured as all cations from an oven-dried soil are extracted with ammonium acetate (A&L labs, 2017; OMAFRA, 2009).

*Table 1. General statistics of the 2017 yield information for the seventeen fields used in this study including crop type for each field. The number of soil samples taken for each field through grid sampling are also incorporated.*

| Field | Crop | Average Yield (bu/acre) | Minimum Yield | Maximum Yield | Yield SD | Number of Yield Observations | Number of Soil Samples |
|---|---|---|---|---|---|---|---|
| Field A | Soybean | 61.08 | 0 | 418.8 | 15.4 | 7897 | 16 |
| Field B | Soybean | 52.97 | 0 | 297.2 | 13.7 | 14063 | 30 |
| Field C | Corn | 236.8 | 0 | 1657.23 | 45.5 | 20069 | 25 |
| Field D | Soybean | 46.87 | 4.6 | 179.3 | 12.7 | 3634 | 5 |
| Field E | Soybean | 51.45 | 5.05 | 188.2 | 11.5 | 4612 | 7 |
| Field F | Soybean | 60.57 | 5.26 | 172.2 | 11.7 | 4885 | 11 |
| Field G | Soybean | 56.46 | 4.84 | 179.5 | 12.1 | 4825 | 10 |
| Field H | Soybean | 58.34 | 4.67 | 225.7 | 12.1 | 5576 | 8 |
| Field I | Soybean | 56.26 | 0 | 1427.2 | 30.1 | 5158 | 8 |
| Field J | Soybean | 51.07 | 0 | 154.1 | 15.6 | 4669 | 7 |
| Field K | Soybean | 44.77 | 0 | 163.39 | 11.9 | 4433 | 7 |
| Field L | Corn | 247.6 | 0 | 592.88 | 62.2 | 10901 | 14 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Field M | Soybean | 231.13 | 0 | 1013.4 | 42.67 | 11143 | 15 |
| Field N | Soybean | 63.3 | 0 | 944.8 | 18 | 9756 | 13 |
| Field O | Corn | 214.89 | 0 | 1708.7 | 50.7 | 15521 | 13 |
| Field P | Corn | 209.88 | 0 | 1623.4 | 60.5 | 13929 | 17 |
| Field Q | Soybean | 43.53 | 0 | 1100.1 | 21.1 | 4423 | 19 |

### 4.2 Data Interpolation

In order to interpolate the soil nutrient attributes to match the same scale as the yield samples, three interpolation methods were used: Thiessen polygons, Kriging, and Inverse Distance Weighting (IDW). Thiessen polygon creates a polygon of influence for each sample and assumes that all values inside the shape are equal (Panagopoulos et al., 2006). Kriging assumes that the distance or direction between sample points reflects spatial correlation used to explain variation in the surface (Chilès and Delfiner, 1999). The IDW interpolator assumed that each input point has a local influence that diminished with distance. It weights the point closer to the processing cell greater than those further away (Longman et al., 1995). The seventeen fields were grouped into five sections to avoid the influence of field edges when the soil points were interpolated. The groups were selected based on which fields were contiguous. The yield point shapefile was then overlaid, and the soil properties were extracted for each crop yield point so that the number of observations was the same between yield and soil nutrients. For each of the interpolation datasets, an MLR, ANN, DT, and RF model were constructed to identify which interpolation dataset was most successful for predicting yield. The models were trained with 70% of each interpolation dataset and tested with the remaining 30%. The points were split into random train and test subset by the Python sklearn.model_selection.train_test_split function. When evaluating each model, the root mean square error, mean absolute error, and R-squared was compared for each

interpolation method. The findings of the analysis are presented in the Appendix and suggested that Kriging, IDW, and Thiessen polygons performed similarly. However, IDW was selected for this study as it had slightly outperformed the other methods.

### 4.3 Topography Data

In addition to soil nutrient data, elevation information was gathered from the Southwestern Ontario Orthophotography Project (SWOOP) 2015 Digital Elevation Model (DEM). SWOOP was created using digital imagery acquired by Fugro using the Leica ADS100 geosystems sensor. The data was collected between April 12th and May 23rd, 2015. The project covers an area of approximately 49,167 $km^2$. Imagery acquisition was performed at 2,377m above mean terrain (AMT) to produce a 20cm full-colour orthorectified image with a horizontal and vertical accuracy of 50cm (SWOOP, 2015). SWOOP is a 2m raster elevation product that serves as a generalized representation of both surface and ground features. The product was generated by an imagery contractor for ortho-rectifying the SWOOP 2015 orthophotography (SWOOP, 2015). Like the soil properties, the yield point shapefiles were overlaid on the elevation raster file, and the elevation variables were extracted for each crop yield point. The elevations across the seventeen fields were quite similar, falling between 187 and 189m above sea level.

The topographic wetness index was obtained for each of the segmented fields using System for Automated Geoscientific Analysis (SAGA GIS). SAGA GIS is an open-source geographic information system that is designed to implement spatial algorithms. The SAGA wetness index was applied to the elevation raster layer to reflect the theoretical distribution of lateral water accumulation (Conrad et al., 2015). The yield

point shapefiles were overlaid on the SAGA wetness index raster file, and the wetness index variables were extracted for each crop yield point.

The yield data were merged with the elevation, wetness index, and soil chemistry data to form a single dataset. Figure 2 provides a spatial representation of the degree of variability present in the scaled crop yield data. The blue areas represent low yield for the soybean and cornfields, red demonstrates high yield levels. Soil and topography spatial variability maps are available in the appendix.

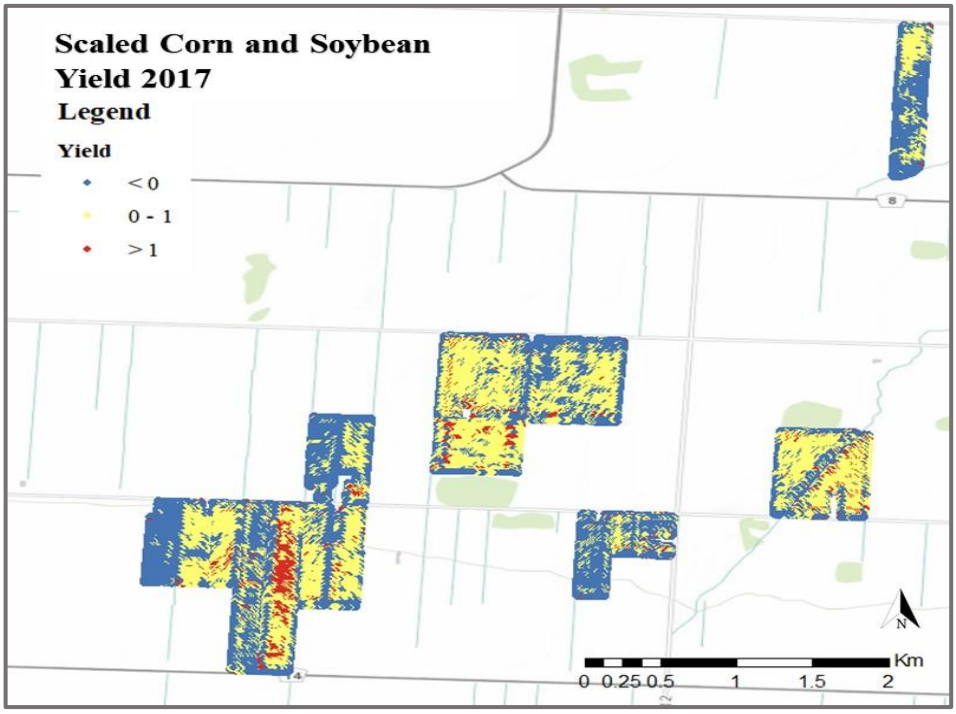



*Figure 2. Spatial variability maps representing the distribution of the scaled corn and soybean yields.*

# 5. Methodology

## 5.1 Variograms

For the estimation of the extent of total sampling and analytical errors, a variographic analysis was carried out in ArcGIS. The semi-variograms were used to determine the ideal scale for sampling yield, soil, and topographic characteristics for the fields of interest. The yield and topographic samples were taken at a much finer resolution than the soil samples. To determine if the spatial distribution of the soil attributes was accommodated, a semi-variogram was developed for each soil attribute. Future sampling processes may be informed by identifying the ideal scale at which samples should be taken for PA applications and understanding the heterogeneity of the samples. Variogram analysis requires the decomposition of variabilities originating from the process and measuring system to decide whether measurements at that scale were able to describe the true process variability with adequate resolution (Engstrom and Esbensen, 2018). In other words, the semi-variogram functions were to quantify the assumption that neighbouring objects appear to be more similar than those farther apart. Furthermore, it measures the strength of the statistical correlation as a distance function (Ebsensen et al., 2015).

The semi-variogram is defined as:

$$\gamma(s_i, s_j) = \frac{1}{2}\, var(Z(s_i) - Z(s_j)), \qquad (1)$$

where var is the variance, $s_i$ and $s_j$ are the points of interest. If $s_i$ and $s_j$ are close to one another in terms of distance, they are likely to be similar than locations far apart and the difference in their values, $Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$, will be small (Engstom and Esbensen, 2018).

The semi-variogram consists of three key parameters: the nugget effect, sill, and range. Each parameter represents a different spatial data variance characteristic. Although theoretically, a variogram should go through the origin (0,0). However, when the variogram does not go through the origin, this is referred to as the nugget effect, which can be interpreted as the minimum practical error (Esbensen et al., 2015). Additionally, the nugget variance may indicate that there are errors during data collection, or samples within a short distance may have substantially different values (Chen et al., 2019). The sill reflects the total variation of the spatial dataset. While the partial sill, the structural variance, represents the intrinsic features of the data. The range is the distance where the variogram reaches the sill, representing the maximum spatial distance at which the dataset can still demonstrate spatial autocorrelation (Chen et al., 2019).

### 5.2 Models and Accuracy Metrics

Four different models were developed in Python and analyzed for this study. The models included multiple linear regression, artificial neural networks, decision trees, and random forests. For the initial comparison, the models were trained with 70% of the data and tested with the remaining 30%. The yield was chosen as the dependent variable, while the soil and topologic variables were selected as the independent variables. The actual and predicted values were compared and evaluated by the following accuracy metrics: relative mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), R-squared, and observed vs. predicted plots. These error measurements are frequently used for agricultural systems and crop models (Jeong et al., 2016).

MAE is the average of differences in estimators (in physical units). It is represented as a percentage relative to the mean yields as yield proportions are different among crops. RMSE measures the difference between the actual and estimates, exaggerating the presence of outliers (Gonzalez-Sanchez et al., 2014; Han & Kamber, 2001). The R-squared value indicates how much of the variance between those two variables can be described by the linear fit. Feature importance was used for the DT, and RF models used to assess the attributes had the most significant effect on the dependent variable. Feature importance calculates each feature's importance as the sum over the number of splits, across all trees that include the feature, proportionally to the number of samples it splits (Altmann et al., 2010).

### *5.2.1 Multiple Linear Regression*

Regression constitutes a supervised learning model, aiming to provide the prediction of output varies according to the input variables. Multiple linear regression has been a popular method for crop yield prediction (Drummond et al., 1995; Drummond et al., 2003; Khakural et al., 1999; Kravchenko & Bullock, 2000). The objective of MLR analysis is to study the relationship between several independent or predictor variables and a dependent or criterion variable (Adamowski et al., 2012). The following equation represents an MLR equation (Pedhazur, 1982):

$$Y = a + \beta_1 X_1 + \cdots + \beta_j X_j, \qquad (2)$$

where Y is a prediction of the response variable, a is the intercept, B is a vector of the slope or coefficients, X is a vector of the predictor variables, and j is the number of predictor variables. For forecasting purposes, the linear regression equation will fit a

forecasting model to an observed data set of Y and X values. The fitted model can be used to forecast the value of Y with new additional observed values of X (Adamowski et al., 2012).

### *5.2.2 Artificial Neural Networks*

An artificial neural network (ANN) can be used to develop empirically-based agronomic models (Kaul et al., 2005). ANNs are inspired by human brain functionality, emulating complex functions such as pattern recognition, cognitive learning, and decision making (Khairunniza-Bejo and Mustaffha, 2014; Gopal and Bhargavi, 2019). This data-driven process captures the relationship between a large number of input and output variables from given patterns. Through these relationships, ANN's can be used to predict future values based on past histories. Commonly, the relationship obtained is non-additive and nonlinear. Thus, nonlinear relationships often overlooked by other techniques can be determined by ANN's with little prior knowledge of the functional relationships (Adamowski et al., 2012; Gopal and Bhargavi, 2019).

The human brain consists of billions of neurons that inter-communicate and process the provided information. Similarly, an ANN consists of interconnected processing units organized in a specific topology. Typically, a minimum of three layers is required in an ANN: the input layer where the data are fed into the system, one or more hidden layers where the learning takes place, and an output layer where the decision/prediction is given (Figure 3). The input and output layers contain nodes that correspond to input and output variables, respectively.

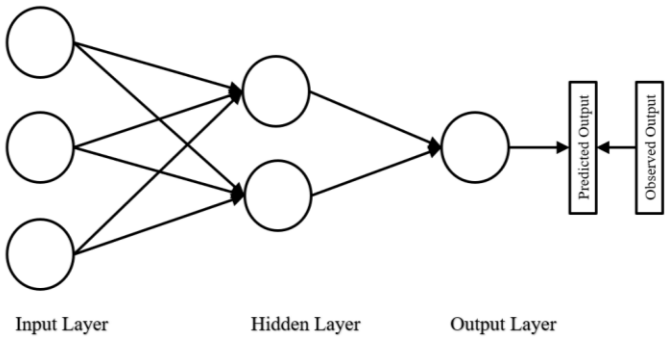



*Figure 3. A simplified illustration of the layers and connections of a three-layer feed-forward back propagating artificial neural network.*

The data moves between layers across weighted interconnections, where the node accepts data from the previous layer and calculates a weighted sum of all its net inputs (Kaul et al., 2005):

$$t_i = \sum_{j=1}^{n} w_{ij} x_j, \qquad (3)$$

Where n is the number of inputs, w is the weight of the connection between node i and j, and x is the input from node j. In order to calculate the node out oi, a transfer function fi, is then applied to the weighted value (Khairunniza-Bejo and Mustaffha, 2014; Kaul et al., 2005).

$$o_i = f(t_i). \qquad (4)$$

The inputs were multiplied by the weights of a node for a given link and summed together. The value is referred to as the node's summed activation. The summed activation is then transformed through an activation function and determines the specific node output. Linear activation is referred to as the simplest activation function, in which no transformation is applied. A network consisting of only linear activation functions is

very simple to train but is often unable to learn complex mapping functions. Nonlinear activation functions are favoured as they allow the nodes to learn more complex data structures (Agarap, 2018). One of the most popular nonlinear activation functions is the sigmoidal function for the hidden and output layers. However, for this study, a rectified linear unit (ReLU) was applied as the activation function as it accounts for the interaction and nonlinear effects of a model. The interaction effect is when one independent variable affects a prediction differently depending on the value of another independent variable (Agarap, 2018).

The ANN model consisted of two hidden layers and used a back-propagation method to train the feed-forward ANN. Back-propagation is often used to minimize potential errors. It is a form of supervised learning where the error rate is sent back through the network to alter the weights to improve prediction, thus decreasing error. However, a large network that uses too many nodes will become over-trained, causing the model to memorize the training data resulting in predictions with higher error. The process is repeated until either a specified error limit is achieved or the total number of training cycles (epochs) has been completed (Gonzalez-Sanchez et al., 2014; Kaul et al., 2005; Khairunniza-Bejo and Mustaffha, 2014; Seyhan et al., 2005).

### *5.2.3 Decision Trees*

The decision tree (DT) algorithm belongs to the supervised learning class, capable of handling both classification and regression-based problems. A DT consists of a flow-chart-like tree structure where various aspects and attributes are considered for evaluating an issue. A recursive algorithm is used for further assessment of classifying the attribute with the highest information (Elavarasan et al., 2018). Furthermore, when a DT is used

for prediction, it is assumed that the response variable's nature is continuous (Raorane and Kulkarni, 2012; Veenadhari, 2011).

A DT model is built from data or observations according to some criteria. The model aims to learn a general rule from the observed instances. (Raorane & Kulkarni, 2012; Veenadhari, 2011). The dataset is gradually organized into small homogenous subsets, while a corresponding tree graph is generated (Liakos et al., 2018). The first node in the tree is named the root node (Gonzalez-Sanchez et al., 2014). Each internal node of the tree structure represents a different pairwise comparison on a selected feature, whereas each branch represents the outcome of this comparison (Liakos et al., 2018). A node with outgoing edges is referred to as a test node, whereas a node without outgoing edges is called a leaf node (Gonzalez-Sanchez et al., 2014). Leaf nodes represent the final decision or prediction after following the root-to-leaf path (expressed as a rule of classification). The variance reduction algorithm was used for continuous target variables (Liakos et al., 2018). This algorithm used the standard formula of variance to choose the best split. The split with lower variance was selected as the criteria to split the population. In this study, the max depth of the tree was 30, which was selected based on the R-squared value. After 30 trees, the model becomes over-trained, resulting in similar or weaker predictions.

#### *5.2.4 Random Forest*

Random Forests (RF) is a non-parametric advanced classification and regression tree (CART) analysis method that consists of multiple decision trees (Jeong et al., 2016). Each tree is built from a random sample of the training data and is drawn with replacements. The data is recursively split into more homogenous units, referred to as

nodes, to improve the response variable's predictability. The most efficient split is defined by identifying the predictor variable and the split point that results in the largest reduction in the residual sum of squares between the sample observations and the node mean. All trees are grown to the maximum extent that is controlled by the size of the nodes. The result is an ensemble of low bias and high variance regression trees, where the final predictions are derived by averaging the predictions of the individual trees (Aghighi et al., 2018; Jeong et al., 2016; Kern et al., 2019).

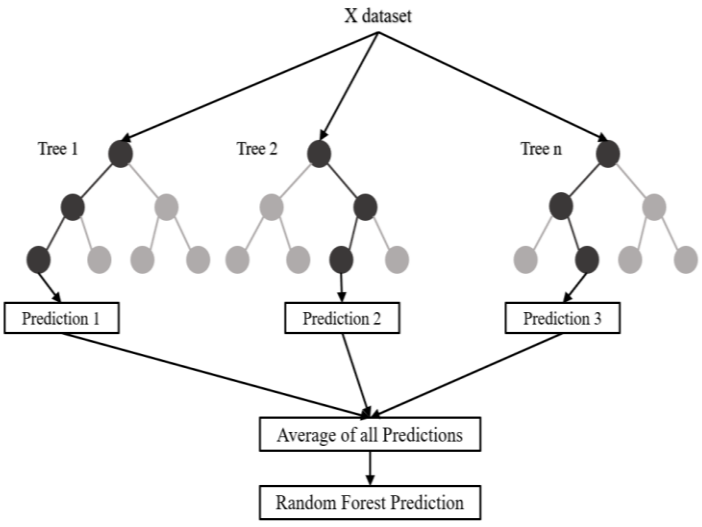



*Figure 4. Representation of the splitting and prediction process of a Random Forest.*

In this study, RF models were built using 30 trees derived from bootstrapped datasets. Bootstrap aggregation attempts to mitigate the problems of high variance and high bias of the final prediction model through a reduction of the correlation between estimators. The minimum number of samples required to be tested at each node was set to the default value of the square root of the total number of predictor variables used. Additionally, the maximum depth of the tree was set to 30, to aid in the comparison between the RF and DT models (Gopal and Bhargavi, 2019; Jeong et al., 2016).

### 5.3 Cross-Validation Techniques

Three different cross-validation methods were used to evaluate the models further. First, a "Jack-Knifing" approach was applied to each field. This approach consisted of eliminating the yield values for one of the fields and predicting the missing yield with the remaining sixteen field values. The "Jack-Knifing" technique used both the DT and RF models to evaluate which model performed better. Additionally, fields that had high error in this analysis were identified as outliers as the surrounding fields were not compatible enough to provide an accurate prediction. Alternatively, there are missing attributes required to predict yield for these fields that were not utilized in this study. For each field, a feature importance calculation was taken to determine the independent variables that had the most significant impact on yield.

Next, a "Leave-Group-Out" cross-validation method was used in the RF model. This method is similar to the previous approach however groups of three fields were selected. The yield data of these three fields were removed, and the remaining fourteen fields datasets were used for training. The groups were selected based on the results from the "Jack-Knifing" approach and consisted of five separate trials comprising of different missing fields. Trial A consisted of the fields that had the lowest R-squared value and highest error in the "Jack-Knifing" analysis, while Trial B was comprised of the top-performing fields. Trial C included three fields that had R-squared values around the median. Finally, Trial D and E were similar in which fields varying in performance were selected.

Finally, a model reduction method was applied to the RF model. This process involved eliminating one attribute at a time based on feature importance from the 70/30

training and testing analysis. The RF model was re-run each time an attribute was removed, and the R-squared and error metrics were compared. The first method of reduction removed attributes with the lowest feature importance values first. Alternatively, the second model of reduction removed the attributes with the highest feature importance values first. The model reduction technique provided insight into which attributes were necessary for yield prediction and was used to cross-validate the feature importance analysis.

# 6. Results

## 6.1 Spatial Structure Analysis

The nugget effect values were small for all the studied attributes. The ranges for elevation, moisture, and yield were longer than the sampling interval of 2m. As well, the range for soil properties CEC, K, P, OM, Zn, and pH were longer than their sampling interval of approximately 110m. Thus, the current sampling designs were enough to reveal the spatial distribution features of these attributes.

In geostatistical theory, the range of the semi-variogram is the maximum distance of the correlated measurements. It can be a sufficient criterion for the selection of sampling design in mapping soil properties (Utset et al., 1998). As a rough guide, the sampling interval should be less than half the variogram range (Kerry and Oliver, 2004; Mallarino et al., 2007). Table 2 provides the nugget, sill, and range values for each attribute. Based on the range of the variograms for CEC, Zn, and yield a suitable sampling interval to ensure reliable kriging estimates would be approximately 300m. K could achieve a suitable sampling interval at around 375m, while OM at 500m, elevation and wetness index at 200m, and finally pH and P at 130m. For these fields, future sampling designs should conform to the current approximate 110m grid sampling design to ensure spatial distribution features are accounted for. Additionally, these results suggest that the sampling intervals were sufficient for modeling.

*Table 2. The nugget, partial sill, and range of the yield, soil and topographic attributes. The range is divided by to identify suitable sampling intervals so that interpolation techniques can identify spatial distribution characteristics.*

| Attribute | Nugget | Partial Sill | Range |
|---|---|---|---|
| Yield | 834 | 7135 | 610 |
| CEC | 5.06 | 4.74 | 664 |
| K | 707 | 1413 | 755 |
| OM | 0.10 | 0.29 | 1017 |
| P | 0.00 | 202 | 263 |
| pH | 0.07 | 0.06 | 283 |
| Zn | 0.54 | 594 | 602 |
| Wetness | 3.17 | 6.44 | 406 |
| Elevation | 0.04 | 0.10 | 471 |

### 6.2 Model Comparison

As previously mentioned, performance metrics were initially used to compare the evaluated techniques. Table 3 shows the results for the MAE, RMSE, and R-squared metrics results for all the field datasets.

*Table 3. Mean absolute error (MAE), root mean square error (RMSE), and R-squared ($r^2$) metrics results for all the evaluated techniques.*

| Method | MAE | RMSE | $r^2$ |
|---|---|---|---|
| MLR | 55.8 | 70.7 | 0.46 |
| ANN | 29.1 | 48.9 | 0.75 |
| DT | 10.5 | 34.2 | 0.90 |
| RF | 10.1 | 27.3 | 0.93 |

Table 4 shows the feature importance of each variable for the RF and DT models. The importance of a feature is the measure of the mean decrease in node impurity, which is computed by the weighted mean squared error of the nodes. Thus, when training a tree, it is possible to determine how much each characteristic reduces the impurity. The impurity of nodes is an indicator of the homogeneity of the labels at the node. Put differently, node impurity is 0 when all patterns at the node are the same. The more impurity a feature decreases, the more significant that feature is. In RF and DT's, each feature's impurity decrease can be averaged across nodes to determine the final

importance of the variables. Features selected at the top of the tree are usually more important than those selected at the end nodes of the trees since the top splits typically lead to greater gains in information (Menze et al., 2009; Sandri and Zucholotto, 2010). In this study, both DT and RF models' feature importance shows that P and pH have the highest values. These results suggest that P and pH are more significant than the other features and lead to bigger information gains within the models.

*Table 4. Feature importance results for both the decision tree (DT) and random forest (RF) models for all the independent variables in relation to yield in the 70/30 training model.*

| Method | Elevation | Wetness | pH | K | OM | CEC | P | ZN |
|---|---|---|---|---|---|---|---|---|
| DT | 0.031 | 0.038 | 0.26 | 0.049 | 0.10 | 0.14 | 0.28 | 0.10 |
| RF | 0.020 | 0.038 | 0.26 | 0.051 | 0.095 | 0.14 | 0.29 | 0.11 |

### 6.3 "Jack-Knifing" Cross-validation

The models of the best performance from the 70/30 training and testing analysis, DT and RF, were utilized in the "Jack-Knifing" cross-validation technique. For the "Jack-Knifing" analysis, one field yields attributes were erased, and the remaining sixteen fields datasets were used to predict the missing yield values. RF and DT models were both used in this analysis as they had the lowest mean error for both metrics and the highest R squared values for all seventeen fields. These results suggest that both RF and DT's are reliable to quantify the relationship between crop yield, soil properties, and topographic characteristics for predicting crop yield in this case. Table 5 shows the "Jack-Knifing" results when all the other fields are used to predict that field.

However, there were three fields that the "Jack-Knifing" approach identified as outliers for both DT and RF models. Field A, Field B, and Field C had the highest mean

error and the lowest R square value. No field identification attributes were included in the model analyses, so Fields A, B, and C were anonymized. Additionally, Fields A, B, and C are contiguous, as presented in Figure 1.

*Table 5. Mean absolute error (MAE), root mean square error (RMSE), and R-squared ($r^2$) metrics results for all the evaluated techniques. The first column shows the dataset identifier. The best result for each field is shown in bold.*

| Field | Method | MAE | MSE | RMSE | $r^2$ | | Method | MAE | MSE | RMSE | $r^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Field A | DT | 63.1 | 8508 | 92.2 | **0.0365** | | RF | 58.8 | 7633 | 87.4 | 0.0327 |
| Field B | DT | 30.0 | 2734 | 52.2 | 0.0182 | | RF | 26.5 | 21828 | 147 | **0.0120** |
| Field C | DT | 64.2 | 10083 | 100 | **0.0287** | | RF | 69.4 | 8940 | 94.5 | 0.0243 |
| Field D | DT | 0.56 | 11.4 | 3.38 | **0.987** | | RF | 5.15 | 89 | 9.45 | 0.904 |
| Field E | DT | 1.16 | 54.0 | 7.35 | **0.974** | | RF | 7.00 | 212 | 14.6 | 0.900 |
| Field F | DT | 2.74 | 36.7 | 6.06 | 0.730 | | RF | 8.31 | 261 | 16.1 | **0.885** |
| Field G | DT | 3.60 | 34.0 | 5.84 | 0.782 | | RF | 8.16 | 274 | 16.6 | **0.871** |
| Field H | DT | 32.5 | 3786 | 61.5 | 0.270 | | RF | 26.0 | 2428 | 49.3 | **0.549** |
| Field I | DT | 23.3 | 2060 | 45.4 | 0.580 | | RF | 18.9 | 1106 | 33.3 | **0.774** |
| Field J | DT | 19.8 | 1883 | 43.4 | 0.592 | | RF | 18.4 | 1818 | 42.6 | **0.606** |
| Field K | DT | 19.0 | 1700 | 41.2 | 0.681 | | RF | 15.3 | 974 | 31.2 | **0.817** |
| Field L | DT | 1.85 | 466 | 21.6 | **0.847** | | RF | 8.85 | 722 | 26.9 | 0.778 |
| Field M | DT | 1.14 | 149 | 12.2 | **0.940** | | RF | 6.61 | 270 | 16.4 | 0.891 |
| Field N | DT | 33.0 | 3970 | 63.0 | 0.228 | | RF | 25.6 | 2216 | 47.1 | **0.569** |
| Field O | DT | 49.2 | 7133 | 84.4 | **0.650** | | RF | 46.6 | 6498 | 80.6 | 0.148 |
| Field P | DT | 3.88 | 835 | 28.9 | **0.784** | | RF | 10.4 | 1029 | 32.1 | 0.734 |
| Field Q | DT | 3.40 | 875 | 29.6 | 0.518 | | RF | 37.4 | 3169 | 56.3 | **0.5267** |

Additionally, through the feature importance analysis, P and pH were selected as the most significant attributes in relevance to yield for the majority of fields. These results further support the feature importance analysis conducted in the 70/30 model results. P and pH had the highest values in the feature importance analysis for Fields A, B, and C in the "Jack-Knifing" examination. However, due to the model's high error, these attributes are not relevant to yield for these fields. Furthermore, Field M feature importance analysis identified CEC and K as the most relevant variables to yield, as shown in Table 6 and Table 7 for both DT and RF models. This difference in feature

importance indicates that Field M would likely need a tailored soil management strategy to compensate for the variable importance discrepancy compared with other fields.

*Table 6. DT variance feature importance for each of the "Jack-Knifing" cross-validation analyses.*

| | Elevation | Wetness | pH | K | OM | CEC | P | Zn |
|---|---|---|---|---|---|---|---|---|
| Field A | 0.122 | 0.033 | 0.263 | 0.054 | 0.041 | 0.125 | 0.371 | 0.100 |
| Field B | 0.012 | 0.033 | 0.263 | 0.054 | 0.041 | 0.125 | 0.371 | 0.100 |
| Field C | 0.039 | 0.054 | 0.078 | 0.039 | 0.112 | 0.454 | 0.040 | 0.183 |
| Field D | 0.036 | 0.031 | 0.251 | 0.056 | 0.066 | 0.150 | 0.299 | 0.111 |
| Field E | 0.036 | 0.032 | 0.251 | 0.056 | 0.061 | 0.159 | 0.295 | 0.110 |
| Field F | 0.042 | 0.034 | 0.246 | 0.045 | 0.064 | 0.156 | 0.302 | 0.111 |
| Field G | 0.037 | 0.035 | 0.260 | 0.043 | 0.069 | 0.149 | 0.303 | 0.105 |
| Field H | 0.047 | 0.036 | 0.246 | 0.052 | 0.067 | 0.136 | 0.310 | 0.107 |
| Field I | 0.047 | 0.035 | 0.293 | 0.048 | 0.073 | 0.114 | 0.032 | 0.077 |
| Field J | 0.055 | 0.035 | 0.240 | 0.053 | 0.065 | 0.149 | 0.304 | 0.099 |
| Field K | 0.042 | 0.037 | 0.244 | 0.051 | 0.066 | 0.144 | 0.306 | 0.110 |
| Field L | 0.042 | 0.028 | 0.265 | 0.049 | 0.064 | 0.156 | 0.278 | 0.117 |
| Field M | 0.019 | 0.040 | 0.138 | 0.240 | 0.083 | 0.277 | 0.079 | 0.123 |
| Field N | 0.057 | 0.036 | 0.235 | 0.048 | 0.064 | 0.144 | 0.305 | 0.111 |
| Field O | 0.034 | 0.038 | 0.227 | 0.045 | 0.061 | 0.157 | 0.313 | 0.125 |
| Field P | 0.046 | 0.031 | 0.263 | 0.055 | 0.065 | 0.163 | 0.267 | 0.109 |
| Field Q | 0.037 | 0.032 | 0.241 | 0.047 | 0.067 | 0.151 | 0.320 | 0.106 |

*Table 7. RF variance feature importance for each of the "Jack-Knifing" cross-validation analyses.*

| | Elevation | Wetness | pH | K | OM | CEC | P | Zn |
|---|---|---|---|---|---|---|---|---|
| Field A | 0.018 | 0.033 | 0.265 | 0.055 | 0.039 | 0.110 | 0.372 | 0.109 |
| Field B | 0.039 | 0.053 | 0.081 | 0.042 | 0.108 | 0.462 | 0.038 | 0.178 |
| Field C | 0.051 | 0.044 | 0.084 | 0.027 | 0.052 | 0.162 | 0.549 | 0.031 |
| Field D | 0.044 | 0.034 | 0.245 | 0.048 | 0.068 | 0.150 | 0.303 | 0.109 |
| Field E | 0.038 | 0.036 | 0.248 | 0.050 | 0.064 | 0.155 | 0.298 | 0.110 |
| Field F | 0.041 | 0.037 | 0.242 | 0.047 | 0.066 | 0.151 | 0.302 | 0.114 |
| Field G | 0.037 | 0.034 | 0.259 | 0.046 | 0.072 | 0.139 | 0.313 | 0.101 |
| Field H | 0.044 | 0.038 | 0.246 | 0.049 | 0.067 | 0.135 | 0.308 | 0.113 |
| Field I | 0.044 | 0.036 | 0.283 | 0.047 | 0.075 | 0.118 | 0.312 | 0.085 |
| Field J | 0.049 | 0.036 | 0.242 | 0.049 | 0.068 | 0.145 | 0.308 | 0.103 |
| Field K | 0.047 | 0.038 | 0.237 | 0.049 | 0.067 | 0.148 | 0.305 | 0.109 |
| Field L | 0.052 | 0.033 | 0.255 | 0.046 | 0.068 | 0.157 | 0.278 | 0.110 |
| Field M | 0.021 | 0.041 | 0.150 | 0.219 | 0.080 | 0.259 | 0.102 | 0.127 |
| Field N | 0.052 | 0.036 | 0.235 | 0.045 | 0.064 | 0.146 | 0.313 | 0.110 |
| Field O | 0.036 | 0.037 | 0.225 | 0.045 | 0.062 | 0.159 | 0.316 | 0.120 |
| Field P | 0.039 | 0.034 | 0.265 | 0.045 | 0.059 | 0.167 | 0.274 | 0.117 |
| Field Q | 0.036 | 0.036 | 0.238 | 0.047 | 0.067 | 0.147 | 0.320 | 0.110 |

### 6.4 “Leave-Group-Out” Cross-validation

The cross-validation technique "Leave-Group-Out" was used to determine how successful the models were in predicting a large amount of missing data. Like the "Jack-Knifing" approach, the "Leave-Group-Out" analysis consisted of erasing the yield data of three fields, while the remaining fourteen fields datasets were used for training. The groups were selected based on the results from the "Jack-Knifing" approach and consisted of five separate trials comprising of different missing fields, as shown in Table 8. With the missing yield data, the RF model predicted approximately 16% missing data for the "Leave-Group-Out" cross-validation method. The RF model was selected for this method as it had slightly outperformed the DT model in the "Jack-Knifing" examination. Trial A consisted of eliminating the yield values for three fields with the highest error in the "Jack-Knifing" analysis. Trial A had the highest RMSE value, and the model explained 6.5% of the yield variation. The poor performance of Trial A suggests that there are factors not accounted for in this study that makes these fields different than the others. As such, fields in Trial A likely cannot be managed the same way as the other fields. Different fields or attributes may need to be introduced to improve the model's performance.

Trial B removed the yield values of the three best-performing fields from the "Jack-Knifing" approach. The predicted yields matched the observed data well with the model explaining 91% of the yield variation. As well, Trial B had the lowest RMSE of the five trials. Trial C erased the yield values of three fields of average performance. The RF model matched the observed data well and explained 89% of the yield variation. The last two Trials included fields that ranged in performance. Trial D and E had acceptable

R-squared values and moderate error values, as shown in Table 8. This cross-validation method demonstrates that even with over 16% missing data if fields share similar characteristics, the RF model can still predict yield. For instance, fields in Trial B, C, D, and E have a similar relationship between soil properties and yield. These similarities suggest that the understanding of Trial D, for example, could be used to manage Trial C as well.

*Table 8. Mean absolute error (MAE), root mean square error (RMSE), and R-squared ($r^2$) metrics results for all the evaluated trials for the "Leave-Group-Out" cross-validation analysis. The identity of the fields that had missing yield values for each trial were also included.*

| | Missing Fields | MAE | RMSE | $r^2$ |
|---|---|---|---|---|
| Trial A | A, B, C | 43.4 | 74.7 | 0.0658 |
| Trial B | D, H, F | 11.1 | 23.6 | 0.917 |
| Trial C | J, E, G | 11.2 | 31.2 | 0.866 |
| Trail D | I, B, D | 20.1 | 39.5 | 0.757 |
| Trial E | Q, I, K | 21.5 | 37.2 | 0.779 |

### 6.5 Model Reduction Cross-Validation

The attribute reduction method assisted in determining the most significant predictors. The ranking of the DT and RF feature importance values were utilized to guide the attribute reduction analysis. The RF model was utilized in this method as it had slightly outperformed the DT in the previous cross-validation methods. Initially, all the attributes were included in the model and had an R squared value of 0.93. The wetness index was removed from the dataset as it had the lowest feature importance value, and the RF model was a rerun. With the wetness index removed, the R squared value only decrease by 0.004. Furthermore, elevation, K, OM, and Zn were removed from the dataset in sequence; however, with the removal of the following attributes, the R-squared

only dropped to 0.91 with relatively low error. These results suggest that CEC, pH, and P account for 91% of the yield variation. Even with only pH and P attributes, which had the highest feature importance values, the model maintained an R-squared of 0.56. The feature importance values provide an indication of which attributes have the highest relevance when predicting yield. The attribute reduction method further supports that pH and P are necessary for crop yield predictions for the fields within this study.

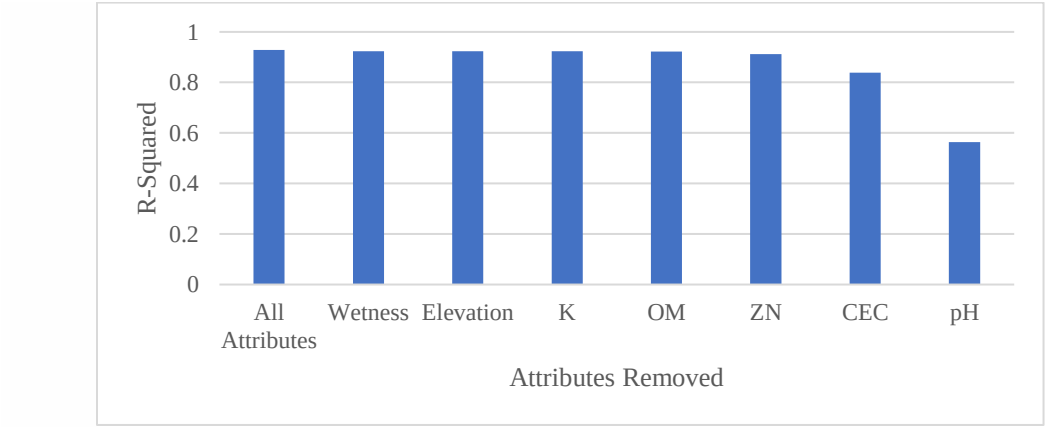



*Figure 5. R-squared values of the first model reduction analysis. The model started with all the attribute. The attributes with the lowest feature importance values were then removed one at a time and the model was re-run after each trial*

A second attribute reduction analysis was conducted; however, this time, the variables with the highest feature importance values were deducted first. For instance, P was the first attribute to removed, and the R-squared value reduced by 0.02. Although the model still had a high R-squared value and low error metrics, this is a sensible decrease compared to the first attribute reduction analysis in which the R-squared value decreased by 0.004. The versatility within the data, as well as the volume of data, compensated for the removal of such a significant attribute. However, with the removal of P, pH, and CEC, the models R-squared dropped to 0.88. Although this is still an acceptable R-squared, the rate at which the R-squared value is dropping, and the increase in error

metrics is quite significant compared to the previous model reduction experiment. When K, elevation, and moisture were the only attributes remaining, the model's R-squared value decreased to 0.73. These results are lower compared to the CEC, pH, and P models results. This assists in supporting that the data's quality is more significant than the quantity of the data.

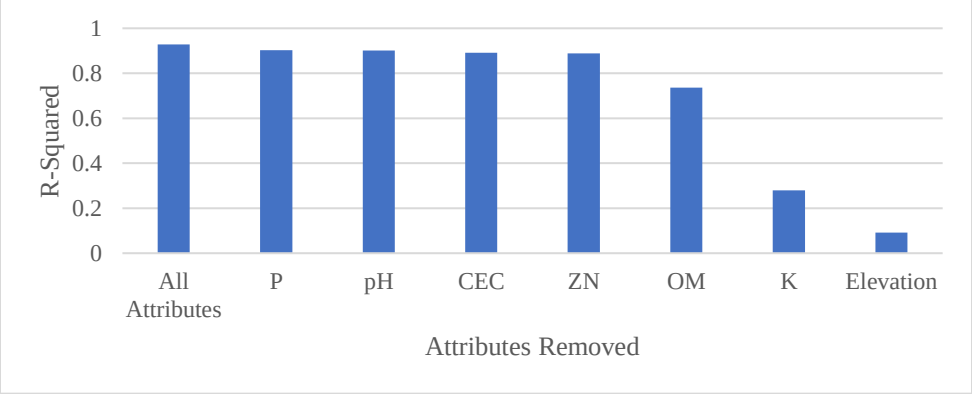



*Figure 6. R-squared values of the first model reduction analysis. The model started with all the attribute. The attributes with the highest feature importance values were then removed one at a time and the model was re-run after each trial*

# 7. Discussion

## 7.1 Models Comparison

The results illustrate that the RF and DT models were useful for crop yield predictions. The RF and DT models outperformed MLR and ANN in the 70/30 testing analysis and the cross-validation methods. While RF and DT have been widely used in ecological studies as a classification algorithm for species distribution and habitat suitability modeling in recent years (Cutler et al., 2007; Lawler et al., 2006), few studies have explored their abilities to regress crop yields or primary productivity studies in agriculture (Jeong et al., 2016). This study demonstrates that RF and DT have many merits which are beneficial for predicting complex crop responses in farming systems and assist in developing farm management strategies.

Most crop models found in pre-precision agriculture literature and during its dawn typically are based on linear regression analysis, calculations of root mean square error, and mean error. MLR techniques using interaction terms are an improvement over strictly linear models (Drummond et al., 1995; Khakural et al., 1999; Kravchenko and Bullock, 2000). MLR and linear mixed models are used in soil mapping. The variability of a target soil property is explained by its relationships among attributes, with shortcomings like autocorrelation and non-linearity between variables (Meersmans et al., 2008). In agricultural practices, a variety of interrelated factors influence crop production. The existence of outliers will complicate MLR's understanding of yield response.

Additionally, MLR models do not provide accurate predictions even in subfield regions considered homogeneous (Drummond et al., 2003; Lambert et al., 2004; Sadler et

al., 2007). The general assumption is that ML techniques are better suited to extract meaningful relationships from data compared to MLR (Gonzalez-Sanchez et al., 2014). MLR is limited by its assumption that fields are homogenous, and the relationships between variables and yield are linear. This study and several others (Khairunniza -Bejo & Mustaffha, 2014; Drummond et al., 2003; Seyhan et al., 2005) found that correlations between yield and soil properties differed considerably within and between fields (Sudduth et al., 1996). There are several benefits of using RF over other methods like conventional MLR when predicting crop yield responses. RF and DT models have been shown to outperform traditional MLR models in explaining data variability (Breiman, 2001; Jeong et al., 2016). There is evidence, for example, that inclusion of extreme temperatures might further improve MLR models (Carlson, 1990; Butler and Huybers, 2013; Schlenker and Roberts; 2009). Climatic data were excluded from this analysis as the primary focus was to identify yield-soil-topological relationships within the fields due to a lack of variability of climate between fields. As well, the fields were exposed to the same environmental conditions over the year. No additional variables to MLR models were added to maintain consistency within all the models. The high performance of RF and DT is apparent when the response is a result of complex interactions between multiple predictors where interactions can complicate modeling.

RF and DT models have an advantage when predictor or explanatory variables are highly correlated. Many variables related to crop production are often strongly correlated with and within each other and may have multicollinearity. Hence, it is reasonable that CEC and pH would be highly correlated (McKenzie et al., 2004). Variable collinearity can be a critical problem in traditional regression models that are derived from linear

regression. RF and DT use the single best variable when splitting responses at each node and averages the predictions of the trees in the forest to make a multidimensional step function. This process suggests that even if multiple variables are correlated and similarly drive the response, only one can affect the RF and DT model at a time (Jeong et al., 2016).

In previous studies, ANN has reported better performance than traditional statistical methods (Jung et al., 2006; Drummond et al., 2003; Sudduth et al., 1997) and regression trees (Ruß and Kruse, 2010; Ruß, 2009). However, in this study, the ANN obtained a high error. The ANN model received the second highest RMSE and MAE values and second-lowest R-squared value. ANN's performance depends on several factors, such as the architecture of the network, training parameters, and samples' reliability. A significant difference between this study and previous ones is that there were no field identification attributes. This study did not include such attributes to prevent site-dependency. Liu at al. (2001) study suggested that the site-dependency of ANNs makes it challenging to achieve adequate results when field identification is missing. Thus, potential upscaling of site-dependent ANN models would be challenging. Additionally, it is impractical to develop different ANN structures for each crop type. Although there are only two crop types in this dataset, if the models were to be applied to other areas, it would be difficult to account for all different crops through this technique (Gonzalez-Sanchez et al., 2014).

As shown in Table 3, both RF and DT models achieved low RMSE and MAE values and a high R-squared. In previous studies, DT and RF are often methods of choice

for prediction as they present hierarchical ranking of feature importance and provides a clear image of active factors (Ebrahimi et al., 2011; Shekoofa et al., 2014). The difference between the RF algorithm and the DT is that the progress of locating the root and dividing the feature nodes takes place at a random phase (Elavarasan et al., 2018). In each of the cross-validation methods, the RF models slightly outperformed DT. The improved performance of RF is likely due to RF models maintain good accuracy despite the presence of outliers and missing data, which is an advantage over DT.

### 7.2 Yield and Topography

In addition to predictive capabilities, RF and DT can also provide useful information about the variable importance and dependence. The rank of feature importance and the partial effect of the variable on the response can be evaluated for systems analysis purposes (Diaz-Uriarte and De Andres, 2006; Jeong et al., 2016; Svetnik et al., 2003; Svetnik et al., 2004). Feature importance and mean decrease accuracy was used to identify the most influential variable determining crop yield in the fields that were tested.

Several studies have suggested that topographic attributes have a significant correlation with yield. For instance, Yang et al. (1998), showed by regression analysis that topographic attributes such as elevation, slope, and aspect have significant correlations with wheat yields. Insight on the spatial-temporal crop variability and its relationship to topographic features is useful for site-specific crop management. It can assist in planning nutrient deposition to compensate for potential erosion and runoff that may impact yield productivity. Furthermore, previous work suggested there is often a negative correlation between yield and elevation. For instance, Changere and Lal (1997),

Kravchenko and Bullock (2000), and McConkey et al. (1997) observed higher yields at lower slope positions and lower yields at high positions. However, in this study, elevation and the topographic wetness index were not relevant to yield. Table 4 presents the correlation values between the topography attributes and yield, and they had the lowest feature importance values of all the variables.

Kravchenko and Bullock (2000) found that a field's slope and curvature largely influenced correlations between yield and topography in different fields. The negative effect of higher topographical location on yield was more intense in fields with a relatively high degree of slope. At the same time, this was less noticeable with fields that have lower slopes. As previously mentioned, Southwestern, Ontario, is relatively flat. The seventeen fields predominately had a slope of 0 degrees, with a maximum slope of 13 degrees. The low slope values are likely to account for the lack of correlation between topography and yield. In the few fields where there is higher variance in slope, the topographic attributes have a more significant impact on yield.

### 7.3 Yield and Soil properties – All Fields

The feature importance analysis revealed P and pH as the most influential variables in the 70/30 training and testing model. As previously discussed, P has a significant yield-limiting factor for annual crop production and plays a crucial role in maintaining a balanced nutrient supply (Fageria, 2001; Robson and Pitman, 1983). P is a major component in plant DNA and RNA and critical in root development, crop maturity, and seed production. The yield maps demonstrate that the majority of the fields are within the optimal P range of 15 to 30 for both corn and soybeans (OMAFRA, 2009). Based on the phosphorus variability map presented in appendix B, there are areas within

the fields below optimal P levels based on the OMAFRA guidelines for corn and soybean fields. In these areas, the crops are likely to depend more on P, potentially limiting the agent for crop production. Thus, additional fertilizer applications are necessary to reach optimum yields.

Additionally, there are areas where P levels are higher than the recommended P range. The use of PA applications may potentially minimize fertilizer applications in these areas. The selective placement of fertilizer would assist in preventing nutrient build-up and potential P-loss from fields. Furthermore, the variability maps in the appendix showed that the higher the spatial variability of P within the fields, the lower the crop yield tends to be. Whereas, fields that have more uniform P levels tend to have higher crop yields.

Soil pH greatly influences nutrient availability. The fertility of soils, generally, decreases with decreasing pH, which can be induced by acidifying nitrogen fertilizer, nitrate leaching, and agricultural practices (McKanzie et al., 2004). Soil pH change can also be caused by natural processes such as decomposition of organic matter and cations' leaching. Southwestern Ontario predominantly has medium to fine-textured mineral soils (Richards, 1949), and the target pH for corn is 6.0 (OMAFRA, 2009). Generally, for cornfields, the yield was higher with higher pH values. However, the pH values of the seventeen fields rarely exceeded 6.5 and predominantly had a pH between 5.8-6.2. Whereas, low yield values for soybean fields tended to occur when the soil's pH was around 6.1-6.3. As previously mentioned, the optimal pH range for soybeans is between 6.6 and 7.0; hence, these low yield values are likely to be influenced by low pH.

Additionally, pH did not exceed 7.0 in these fields, so there are no areas within this study that at the chance of the soil being too basic.

Regarding the distribution of other nutrients within the fields, OM followed a similar spatial distribution pattern as P. Areas high in P, for example, also had greater quantities of OM. This similarity is likely due to the fertilizer, which tends to have high concentrations of both OM and P. As well, OM significantly influences P as the rate of OM decomposition influences the rate at which P is released. Furthermore, in the Zn variability map, areas high in Zn follow a similar spatial distribution to areas high in P; alternatively, areas low in P do not necessarily have low Zn concentrations. The similar spatial distribution of Zn is potentially due to Zn being added as a fertilizer.

The results of the 70/30 feature importance analysis were questioned by the model reduction cross-validation method as it suggested that there is some redundancy in the independent variables. This is to be anticipated, as, in the 70/30 feature importance analysis, OM, Zn, CEC, and K were all relevant to yield. All these soil properties played a significant role in predicting yield and follow similar spatial distribution within the fields as these variables are somewhat intertwined, which complicated the interpretation. So, while K and OM are likely less important than P and pH, it is difficult to say that the fields are only P and pH dependent. Instead, Zn, OM, elevation, and wetness index account for 88% of yield variation. All these variables are likely to co-vary with P, and the soils may tend to retain more P. This is not to suggest that P is not important to yield; in the "Jack-Knifing" function value analysis, it provided the most information gained for almost all the fields.

### 7.4 Yield and Soil Properties – Individual Fields

Through the "Jack-Knifing" cross-validation technique, a feature importance analysis was conducted on each field. All of the fields except for two found P and pH as the most important features, which are consistent with the 70/30 analysis feature importance conducted in the. Field B and Field M found K and CEC had the most considerable influence on yield. However, Fields A, B, C, and N all performed poorly with high error in the "Jack-Knifing" cross-validation assessment. These results suggested that the features selected for these fields are likely not significant. There are potentially additional attributes that were not considered in this study, which are related to these fields yield, such as farm management practices or additional soil attributes. By identifying the attributes correlated with these distinctive fields, the performance of the model could be improved. Additional years of yield could also be added to the study to add temporal diversity, which may assist in predicting yield for these fields.

Unlike Field B, Field M performed well in the "Jack-Knifing" assessment. The high performance in the cross-validation technique suggests that Field M is likely dependent on K and CEC. K is an essential nutrient for crop development and plays a variety of roles in plant metabolism processes (Dibb and Thompson, 1985). Potassium has several physiological functions in plant cells, such as enzyme activation and balancing the charge of anions (Fageria, 2001). Adequate K level is essential for the efficient use of N in crop plants. However, Field M and its adjoining Field L K values are less than 100ppm, as shown in the appendix's K variability map. K values below 120pm

are considered below optimal levels for corn and soybean fields. Additional, K applications are likely necessary to accommodate Field M.

### 7.5 Predicting Missing Data for Multiple Fields

The "Leave-Group-Out" cross-validation methods suggest that fields of similar soil and topographic characteristics can be used to assist in predicting missing yield. This analysis can be useful for predicting yield for a field that has not yet been sampled or has missing data. For instance, with 16% of the overall dataset missing the RF model was successful at predicting yield at a relatively high level of accuracy. However, for the model to perform well, the fields need to have similar topologic and soil characteristics as the surrounding fields. In Trial A, when the outlier fields were selected as the three missing fields, the predictive model performed poorly with high error and a low R-squared. The success of the model depends on the similarity between the fields and the variability of the data. As well, the fields were exposed to similar farm management strategies and climatic conditions. However, if the data are too similar, the model will not have the training data to compensate for potential outliers. As PA continues to expand, and additional variables and potential fields are introduced, the ability to predict missing field values is likely to improve.

For this study, fields were grouped as they were contiguous; no additional analysis was conducted to determine how the fields should be grouped. However, additional research, such as a "Jack-Knifing" analysis, may help to classify which fields should be grouped if the fields are all contiguous. By splitting the fields into similar

classes, users can apply the understanding of one group of fields to another. Hence, the "Leave-Group-Out" indicates that there might be a compelling case to apply what is learned in some fields to others through PA, but the user needs to know which fields' group' together.

### 7.6 Sampling Points

Previous precision agricultural studies with ML are often constrained by the restricted number of fields and soil sample points (Jung et al., 2006; Drummond et al., 2003). For instance, approximately 90% of the PA studies recorded in the International Precision Agriculture Conference between 1999 to 2004, were conducted in single fields on commercial farms (McBratney et al., 2005). Most work that considers several fields, however, are mainly conducted on different farms. The challenge of PA is to become an integral part of the normal farming process. It is, therefore, ideal that all fields on a farm are monitored so that PA practices can be applied. This study provides a fair representation of multiple fields being monitored while undergoing similar management practices. This provides a strong foundation for PA procedures to be applied once spatial variability of soil nutrients and topographic attributes has been analyzed.

A diverse training dataset for ML models is crucial to achieving optimal results. A dataset that lacks diversity will result in overfitting, which occurs when a feature is too closely related to a limited set of observations (Sarvari, 2010). Datasets from previous studies often include several years of data for a select number of fields. As described above, Drummond et al. (2003) compared the ML and regression applications and

defined ANNs as the most effective method for predicting crop yields. They collected climatic data and soil properties from three fields ranging from 13-36 ha in size from 1993-1997. In total, they utilized 3120 observations in their study. Such research takes into account the temporal yield variability across a field, which is necessary to account for the year-to-year variations. However, Drummond et al. (2003) study encompass a relatively small number of fields that restricts spatial diversity, especially if the field properties are similar.

This study focussed on a single year event and did not account for the temporal variability of the fields. Nevertheless, the lack of temporal diversity was compensated by the spatial variability of the seventeen fields. Additionally, the geostatistical analysis from the semi-variograms demonstrated that the current sampling design was good enough to reveal the spatial distribution of the yield, soil, and topographic attributes. Hence, when the soil attributes were interpolated to match the yield and topographic attributes sample points, the spatial distribution of the variables was accounted for. The semi-variograms also provided an outline to aid with future soil sampling designs as it provided insight into a suitable sampling interval for each attribute to warrant reliable kriging estimates. For example, fields should be sampled at a fine enough resolution to account for attributes of P and pH, as they have the most significant effect on yield and require the finest sampling resolution.

# 8. Conclusion

This study evaluated MLR and ML techniques' effectiveness in modeling complex yield responses of corn and soybean crops in seventeen fields. In the 70/30 analysis, RF and DT models outperformed the other ML and regression algorithms in predicting yield. The cross-validation techniques took the comparative analysis a step further. They identified which attributes had the most significant impact on yield, what attributes are necessary for crop yield prediction, and predicted yield with up to 16% missing data. Based on these analyses, it has been shown that soil nutrients are useful in describing yield variability on an agricultural field scale. The soil characteristics were particularly useful in site-specific management for delineating areas. Topographic properties did not play a significant role in predicting crop yield. Nevertheless, topographic data in fields with more significant variation in the elevation and wetness index are likely to have a more notable effect on yield.

Through the model reduction cross-validation analysis, it was determined that although P and pH had the most significant influence on yield for most of the feature importance analysis, the soil properties are likely correlated. They are thus complicating the interpretation. So, while the other soil properties are probably less important than P and pH, it does not necessarily mean the fields depend only on P and pH. However, the cross-validation methods were effective at identifying outlier fields. For instance, Field M was identified as an outlier due to its low K concentrations, especially compared to other fields. Customized farm management plans should be developed for outlier fields to better account for the soil-topographic differences.

This study's result presents the potential for the implementation of the RF and DT algorithms to assist with farm management practices. However, the performance of the model varies from field to field. In some cases, this information is capable of explaining a substantial portion of the yield variability. In contrast, in other cases, only a small portion of the yield variability can be explained. Finally, before applying ML algorithms, geostatistical analyses should be conducted to ensure the sampling methods account for the spatial distribution of each attribute. In summary, the results support that RF, DT, and cross-validation techniques can be useful for predicting crop yield and for farm management practices.

# Appendix A. Interpolation Methods Comparison

*Table A1. Mean absolute error (MAE), root mean square error (RMSE), and R-squared ($r^2$) metrics results for Theisen polygon interpolation.*

| Method | MAE | MSE | RMSE | $r^2$ |
|---|---|---|---|---|
| MLR | 62.3 | 5959 | 77.1 | 0.35 |
| ANN | 33.4 | 2421 | 49.2 | 0.73 |
| DT | 18.2 | 1540 | 39.2 | 0.83 |
| RF | 17.6 | 1451 | 38.1 | 0.84 |

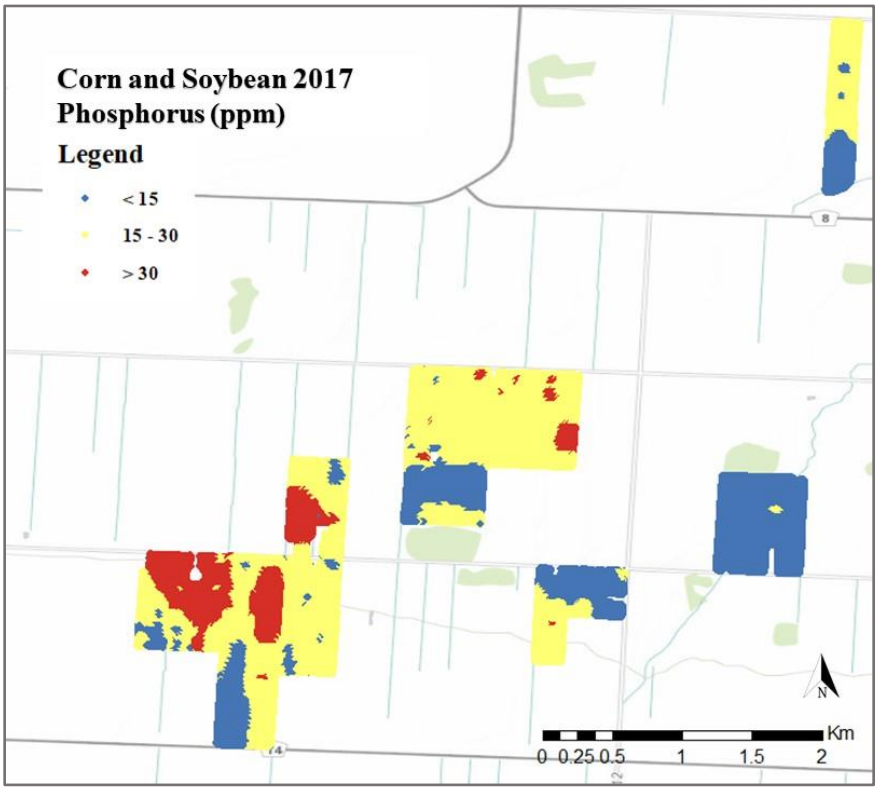
*Table A2. Mean absolute error (MAE), root mean square error (RMSE), and R-squared ($r^2$) metrics results for Kriging interpolation.*

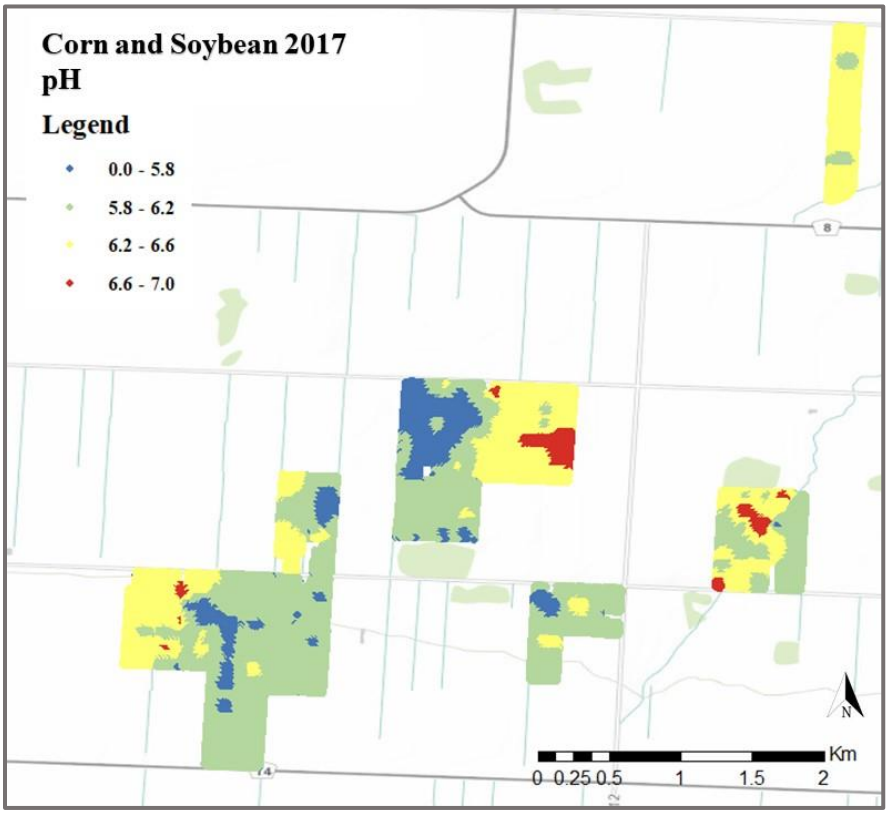
| Method | MAE | MSE | RMSE | $r^2$ |
|---|---|---|---|---|
| MLR | 51.7 | 4449 | 66.7 | 0.51 |
| ANN | 27.1 | 2278 | 45.7 | 0.70 |
| DT | 12.5 | 1220 | 34.9 | 0.86 |
| RF | 10.5 | 876 | 29.6 | 0.90 |

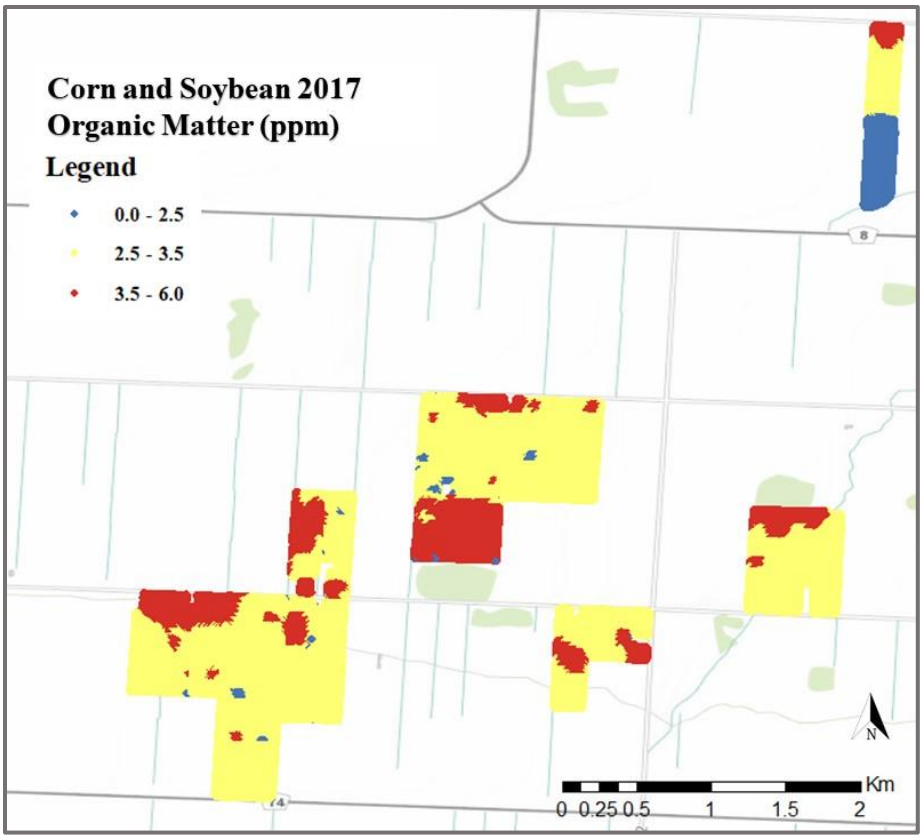
*Table A3. Mean absolute error (MAE), root mean square error (RMSE), and R-squared ($r^2$) metrics results for IDW interpolation.*

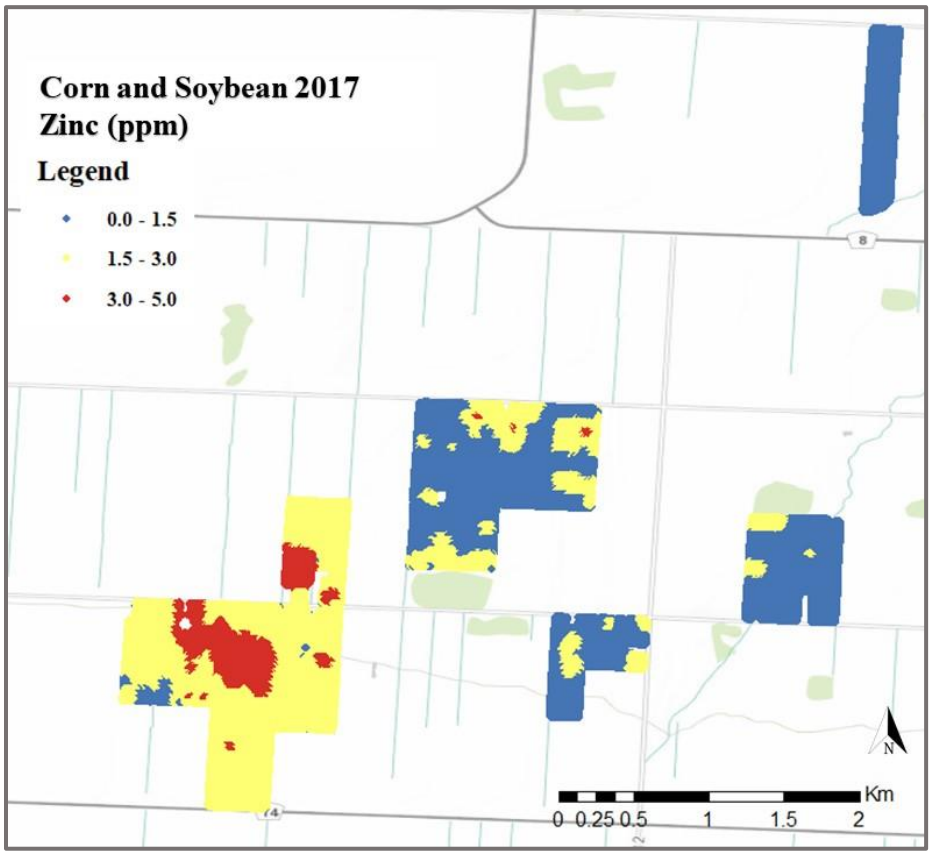
| Method | MAE | MSE | RMSE | $r^2$ |
|---|---|---|---|---|
| MLR | 55.81 | 5000 | 70.71 | 0.46 |
| ANN | 29.13 | 2389 | 48.88 | 0.75 |
| DT | 10.51 | 973 | 34.19 | 0.89 |
| RF | 10.07 | 746 | 27.31 | 0.93 |

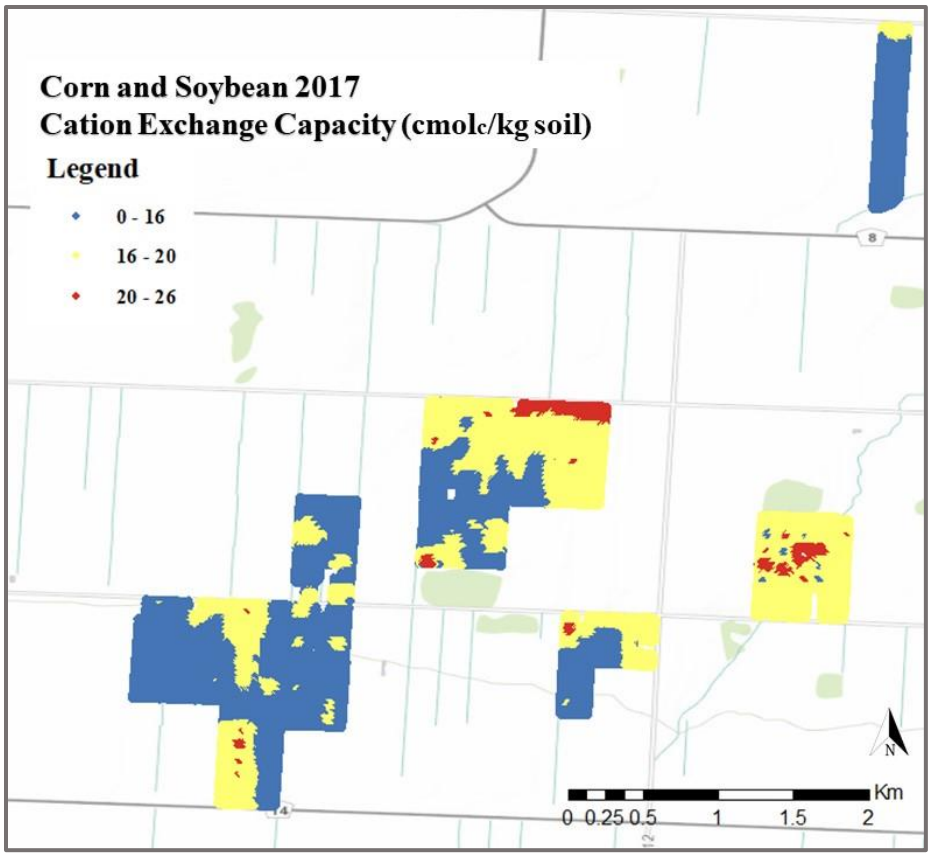
# Appendix B. Soil and Topography Maps

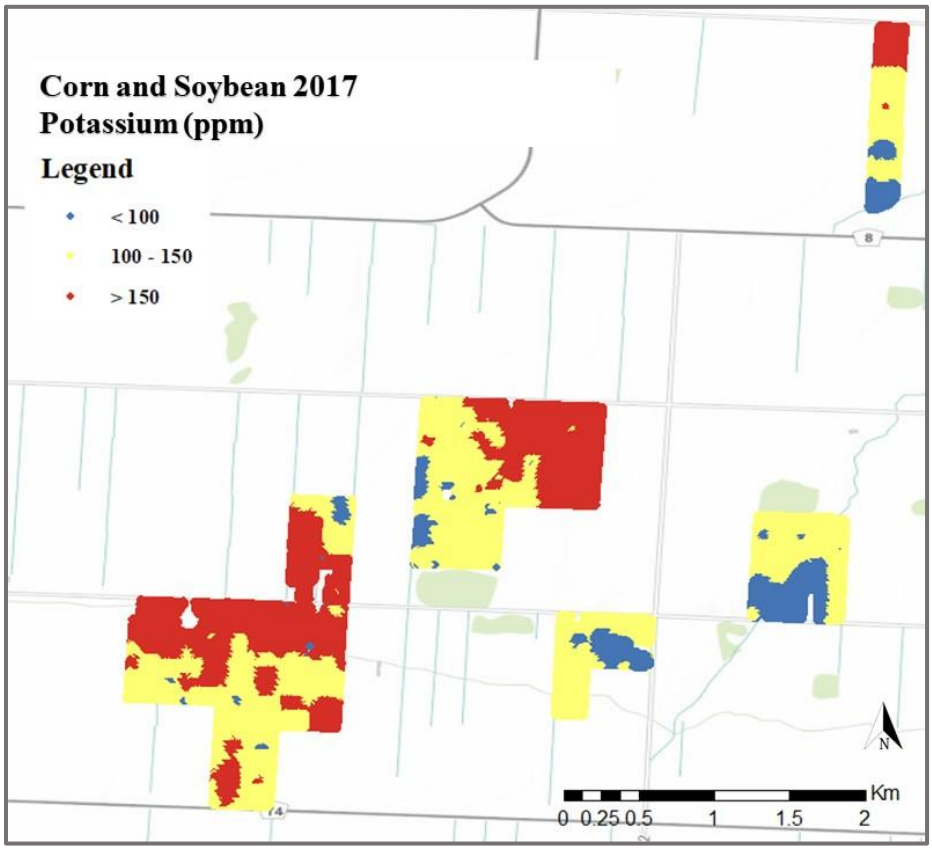



Corn and Soybean 2017
Organic Matter (ppm)
Legend
0.0 - 2.5
2.5 - 3.5
3.5 - 6.0
N
Km
0 0.25 0.5 1 1.5 2


Corn and Soybean 2017
Zinc (ppm)
Legend
0.0 - 1.5
1.5 - 3.0
3.0 - 5.0
N
Km
0 0.25 0.5 1 1.5 2


Corn and Soybean 2017
Cation Exchange Capacity (cmolc/kg soil)
Legend
0 - 16
16 - 20
20 - 26
N
Km
0 0.25 0.5 1 1.5 2


Corn and Soybean 2017
Potassium (ppm)
Legend
< 100
100 - 150
> 150
N
Km
0 0.25 0.5 1 1.5 2

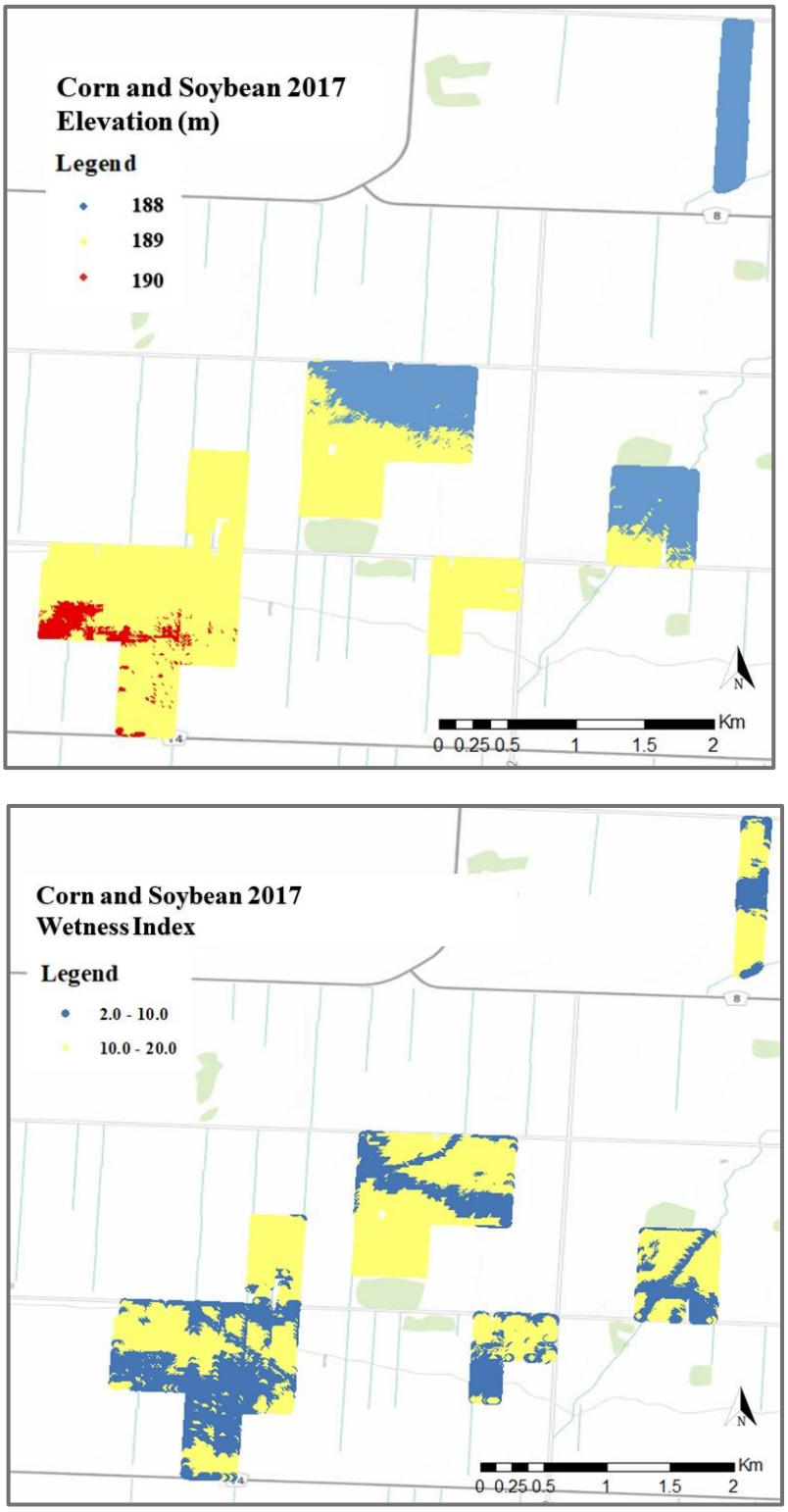



*Figure A1. Spatial variability maps representing the distribution of soil and topographic properties of the seventeen fields. The optimal concentrations were determined following the OMAFRA (2009) corn and soybean recommendations. A smaller number of bins were used to identify areas of low, high, and optimal values so that it would be easier to compare the fields.*

# References


Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B., & Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research*, *48*(1).

Afyuni, M. M., Cassel, D. K., & Robarge, W. P. (1993). Effect of landscape position on soil water and corn silage yield. *Soil Science Society of America Journal*, *57*(6), 1573-1580.

Agarap, A. F. (2019). Deep learning using rectified linear units (ReLU).

Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H. S., & Radiom, S. (2018). Machine learning regression techniques for the silage maize yield prediction using time-series images of Landsat 8 OLI. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(12), 4563-4577.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., & Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing*, *7*(12), 16398-16421.

Alloway, B. J. (2001). Zinc-the vital micronutrient for healthy, high-value crops. *International Zinc Association, Brussels*.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340-1347.

Asare, E., & Segarra, E. (2018). Adoption and extent of adoption of georeferenced grid soil sampling technology by cotton producers in the southern US. *Precision agriculture*, *19*(6), 992-1010.

A&L Labs. (2011). *Soil Analysis Reference Guide*. 6–9. Retrieved from http://www.alcanada.com/index_htm_files/Soil_Analysis_Guide.pdf

Bakhsh, A., Colvin, T. S., Jaynes, D. B., Kanwar, R. S., & Tim, U. S. (2000). Using soil attributes and GIS for interpretation of spatial variability in yield. *Transactions of the ASAE*, *43*(4), 819.

Basso, B., Cammarano, D., & Carfagna, E. (2013, July). Review of crop yield forecasting methods and early warning systems. In *Proceedings of the first meeting of the scientific advisory committee of the global strategy to improve agricultural and rural statistics, FAO Headquarters, Rome, Italy* (Vol. 41).

Bejo, S. K., & Mustaffha, S. (2014). *Application of Artificial Neural Network in Predicting Crop Yield: A Review*.

Bendre, M. R., Thool, R. C., & Thool, V. R. (2015, September). Big data in precision agriculture: Weather forecasting for future farming. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)* (pp. 744-750). IEEE.

Bergstrom, W. G., Cox, W. J., Ferguson, G. A., Klausner, S. D., Pardee, W. D., Reid, W. S., ... & Wright, M. J. (1987). Cornell field crops and soils handbook.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Butler, E. E., & Huybers, P. (2013). Adaptation of US maize to temperature variations. *Nature Climate Change*, *3*(1), 68-72.

Carlson, R. E. (1990). Heat stress, plant-available soil moisture, and corn yields in Iowa: A short-and long-term view. *Journal of Production Agriculture*, *3*(3), 293-297.

Castle, M. H., Lubben, B. D., & Luck, J. D. (2016). Factors influencing the adoption of precision agriculture technologies by Nebraska producers.

Chang, H. H., Mishra, A. K., & Livingston, M. (2011). Agricultural policy and its impact on fuel usage: Empirical evidence from farm household analysis. *Applied energy*, *88*(1), 348-353.

Changere, A., & Lal, R. (1997). Slope position and erosional effects on soil properties and corn production on a Miamian soil in central Ohio. *Journal of Sustainable Agriculture*, *11*(1), 5-21.

Chen, L., Gao, Y., Di Zhu, Y. Y., & Liu, Y. (2019). Quantifying the scale effect in geospatial big data using semi-variograms. *PloS one*, *14*(11).

Chilès, J. P., & Delfiner, P. (1999). Geostatistics: Modeling Spatial Uncertainty, A John Wiley & Sons. *Inc., Publication, ISBN-13*, 978-0.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, 61-69.

Ciha, A. J. (1984). Slope Position and Grain Yield of Soft White Winter Wheat 1. *Agronomy Journal*, *76*(2), 193-196.

Coble, K., Griffin, T., Ahearn, M., Ferrell, S., McFadden, J., Sonka, S., & Fulton, J. (2016). *Advancing US agricultural competitiveness with big data and agricultural economic market information, analysis, and research* (No. 643-2016-44464).

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., ... & Böhner, J. (2015). System for automated geoscientific analyses (SAGA) v. 2.1. 4. *Geoscientific Model Development Discussions*, *8*(2).

Cook, S. E., & Bramley, R. G. V. (1998). Precision agriculture—opportunities, benefits and pitfalls of site-specific crop management in Australia. *Australian Journal of Experimental Agriculture*, *38*(7), 753-763.

Cook, S. E., & Bramley, R. G. V. (2001). Is agronomy being left behind by precision agriculture. In *Proceedings of the 10th Australian Agronomy Conference'. Hobart, Tas.(The Australian Society of Agronomy)*.

Corwin, D. L., & Lesch, S. M. (2003). Application of soil electrical conductivity to precision agriculture. *Agronomy journal*, *95*(3), 455-471.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

Dahikar, S. S., & Rode, D. S. V. (2014). Agricultural Crop Yield Prediction Using Artificial Neural Network Approach" published in International Journal Of Innovative Research In Electrical. *Electronics, Instrumentation And Control Engineering*, *2*(1).

D'Amario, S. C., Rearick, D. C., Fasching, C., Kembel, S. W., Porter-Goff, E., Spooner, D. E., ... & Xenopoulos, M. A. (2019). The prevalence of nonlinearity and detection of ecological breakpoints across a land use gradient in streams. *Scientific reports*, *9*(1), 1-11.

Daniels, R. B., Gilliam, J. W., Cassel, D. K., & Nelson, L. A. (1987). Quantifying the effects of past soil erosion on present soil productivity. *Journal of Soil and Water Conservation*, *42*(3), 183-187.

D'Antoni, J. M., Mishra, A. K., & Joo, H. (2012). Farmers' perception of precision technology: The case of autosteer adoption by cotton farmers. *Computers and Electronics in agriculture*, *87*, 121-128.

Da Silva, J. M., & Silva, L. L. (2008). Evaluation of the relationship between maize yield spatial and temporal variability and different topographic attributes. *Biosystems engineering*, *101*(2), 183-190.

Delmotte, S., Tittonell, P., Mouret, J. C., Hammond, R., & Lopez-Ridaura, S. (2011). On farm assessment of rice yield variability and productivity gaps between organic and conventional cropping systems under mediterranean climate. European Journal of Agronomy, 35(4), 223–236.

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, *7*(1), 3.

Dibb, D. W., & Thompson Jr, W. R. (1985). Interaction of potassium with other nutrients. *Potassium in agriculture*, 515-533.

Drummond, S. T., Sudduth, K. A., & Birrell, S. J. (1995). Analysis of spatial factors influencing crop yield. *Precision Agriculture*, (precisionagricu3), 129-139.

Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J., & Kitchen, N. R. (2003). Statistical and neural methods for site–specific yield prediction. *Transactions of the ASAE*, *46*(1), 5.

Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimie, E., & Ebrahimi, M. (2011). Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. *PloS one*, *6*(8).

Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., & Srinivasan, K. (2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Computers and electronics in agriculture*, *155*, 257-282.

Elrashidi, M. A. (2010). Selection of an appropriate phosphorus test for soils. *Testing methods for phosphors and organic matter*.

Engström, K., & Esbensen, K. H. (2018). Variographic Assessment of Total Process Measurement System Performance for a Complete Ore-to-Shipping Value Chain. *Minerals*, *8*(7), 310.

Erickson, B., Lowenberg-DeBoer, J., & Bradford, J. (2017). 2017 Precision agriculture dealership survey. *Purdue University*.

Erickson, B., & Widmar, D. A. (2015). Precision agricultural services dealership survey results. *Purdue University. West Lafayette, Indiana, USA*, 37.

Esbensen, K. H., & Romañach, R. J. (2015, June). Proper sampling, total measurement uncertainty, variographic analysis & fit-for-purpose acceptance levels for pharmaceutical mixing monitoring. In *TOS forum* (Vol. 5, pp. 25-30).

Fageria, V. D. (2001). Nutrient interactions in crop plants. *Journal of plant nutrition*, *24*(8), 1269-1290.

Fageria, N. K., Baligar, V. C., & Li, Y. (2006). Enhancing phosphorus use efficiency in crop plants grown on Brazilian oxisols. *VMC Alves et al*, 14-19.

Fiez, T. E., Miller, B. C., & Pan, W. L. (1994). Winter wheat yield and grain protein across varied landscape positions. *Agronomy Journal*, *86*(6), 1026-1032.

Fleming, K. L., Westfall, D. G., Wiens, D. W., & Brodahl, M. C. (2000). Evaluating farmer defined management zone maps for variable rate fertilizer application. *Precision Agriculture*, *2*(2), 201-215.

Fortin, J. G., Anctil, F., Parent, L. É., & Bolinder, M. A. (2011). Site-specific early season potato yield forecast by neural network in Eastern Canada. *Precision agriculture*, *12*(6), 905-923.

Frank, R., & Ripley, B. D. (1977). Land use activities in eleven agricultural watersheds in Southern Ontario, Canada, 1975-76.Franzen, D. W., & Peck, T. R. (1995). Field soil sampling density for variable rate fertilization. *Journal of Production Agriculture*, *8*(4), 568-574.

Frausto-Solis, J., Gonzalez-Sanchez, A., & Larre, M. (2009, November). A new method for optimal cropping pattern. In *Mexican International Conference on Artificial Intelligence* (pp. 566-577). Springer, Berlin, Heidelberg.

Gburek, W. J., & Sharpley, A. N. (1998). Hydrologic controls on phosphorus loss from upland agricultural watersheds. *Journal of Environmental Quality*, *27*(2), 267-277.

Gopal, P. M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. *Computers and Electronics in Agriculture*, *165*, 104968.

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, *12*(2), 313–328.

Grace, P. R., Robertson, G. P., Millar, N., Colunga-Garcia, M., Basso, B., Gage, S. H., & Hoben, J. (2011). The contribution of maize cropping in the Midwest USA to global warming: A regional estimate. *Agricultural Systems*, *104*(3), 292-296.

Han, J., & Kamber, M. (2001). Data Mining Concepts and Techniques, Morgan Kaufmann Publishers. *San Francisco, CA*, 335-391.

Hanna, A. Y., Harlan, P. W., & Lewis, D. (1982). Soil available water as influenced by landscape position and aspect 1. *Agronomy Journal*, *74*(6), 999-1004.

Hazelton, P., & Murphy, B. (2007). *Interpreting soil test results: What do all the numbers mean?*. CSIRO publishing.

Hodgson, J. F., Lindsay, W. L., & Trierweiler, J. F. (1966). Micronutrient cation complexing in soil solution: II. Complexing of zinc and copper in displaced solution from calcareous soils. *Soil Science Society of America Journal*, *30*(6), 723-726.

Holt, R. F., Timmons, D. R., Voorhees, W. B., & Van Doren, C. A. (1964). Importance of Stored Soil Moisture to the Growth of Corn in the Dry to Moist Subhumid Climatic Zone 1. *Agronomy Journal*, *56*(1), 82-85.

Hoogenboom, G., White, J. W., & Messina, C. D. (2004). From genome to crop: integration through simulation modeling. *Field Crops Research*, *90*(1), 145-163.

Iqbal, J., Read, J. J., Thomasson, A. J., & Jenkins, J. N. (2005). Relationships between soil–landscape and dryland cotton lint yield. *Soil Science Society of America Journal*, *69*(3), 872-882.

Irmak, A., J. W. Jones, W. D. Batchelor, and J. O. Paz. (2001) Estimating spatially variable soil properties for application of crop models in precision farming. *Transactions of the ASAE* 44, no. 5 (2001): 1343.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PLoS One*, *11*(6).

Jung, W. K., Kitchen, N. R., Sudduth, K. A., & Anderson, S. H. (2006). Spatial characteristics of claypan soil properties in an agricultural field. *Soil Science Society of America Journal*, *70*(4), 1387-1397.

Jochinke, D. C., Noonon, B. J., Wachsmann, N. G., & Norton, R. M. (2007). The adoption of precision agriculture in an Australian broadacre cropping system—Challenges and opportunities. *Field Crops Research, 104*(1-3), 68-76.

Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems, 85*(1), 1-18.

Kaspar, T. C., Colvin, T. S., Jaynes, D. B., Karlen, D. L., James, D. E., Meek, D. W., ... & Butler, H. (2003). Relationship between six years of corn yields and terrain attributes. *Precision agriculture*, *4*(1), 87-101.

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. In *Survey Research Methods* (Vol. 13, No. 1, pp. 73-93).

Kerry, R., & Oliver, M. A. (2004). Average variograms to guide soil sampling. *International Journal of Applied Earth Observation and Geoinformation, 5*(4), 307-325.

Ketterings, Q. M., Czymmek, K. J., & Klausner, S. D. (2003). Phosphorus guidelines for field crops in New York. *Second Release. Department of Crop and Soil Sciences Extension Series E03-15. Ithaca, NY: Cornell University*.

Ketterings, Q. M., Klausner, S. D., & Czymmek, K. J. (2003). Nitrogen guidelines for field crops in New York. *Second Release. Department of Crop and Soil Extension Series E03-16. Cornell University, Ithaca, NY*.

Ketterings, Q. M., Klausner, S. D., & Czymmek, K. J. (2003). Potassium guidelines for field crops in New York. *Ext. Ser. E03–14. Dep. of Crop Soil Sci. Cornell Univ., Ithaca, NY*.

Ketterings, Q., Reid, S., & Rao, R. (2007). Cation exchange capacity (CEC). *Fact sheet*, *22*.

Khairunniza-Bejo, S., Mustaffha, S., & Ismail, W. I. W. (2014). Application of artificial neural network in predicting crop yield: A review. *Journal of Food Science and Engineering*, *4*(1), 1.

Khakural, B. R., Robert, P. C., & Huggins, D. R. (1999). Variability of corn/soybean yield and soil/landscape properties across a southwestern Minnesota landscape. *Precision agriculture*, (precisionagric4a), 573-579.

Khazaei, J., Naghavi, M. R., Jahansouz, M. R., & Salimi-Khorshidi, G. (2008). Yield estimation and clustering of chickpea genotypes using soft computing techniques. *Agronomy Journal*, *100*(4), 1077–1087.

Kim, N., & Lee, Y. W. (2016). Machine learning approaches to corn yield estimation using satellite images and climate data: A case of Iowa State. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, *34*(4), 383-390.

Kitchen, N. R., Sudduth, K. A., & Drummond, S. T. (1999). Soil electrical conductivity as a crop productivity measure for claypan soils. *Journal of Production Agriculture*, *12*(4), 607-617.

Kravchenko, A. N., & Bullock, D. G. (2000). Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal*, *92*(1), 75-83.

Krishnan, M. (2019). Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology*, 1-16.

Lacy, J. (2011). Cropcheck: Farmer benchmarking participatory model to improve productivity. *Agricultural Systems*, *104*(7), 562-571.

Lahoche, F., Godard, C., Fourty, T., Lelandais, V., & Lepoutre, D. (2002, July). An innovative approach based on neural networks for predicting soil component variability. In *Proceedings of the 6th International Conference on Precision Agriculture and Other Precision Resources Management, Minneapolis, MN, USA* (pp. 14-17).

Lambert, D. M., Lowenberg-Deboer, J., & Bongiovanni, R. (2004). A comparison of four spatial regression models for yield monitor data: A case study from Argentina. *Precision Agriculture*, *5*(6), 579-600.

Landau, S., Mitchell, R. A. C., Barnett, V., Colls, J. J., Craigon, J., & Payne, R. W. (2000). A parsimonious, multiple-regression model of wheat yield response to environment. *Agricultural and forest meteorology*, *101*(2-3), 151-166.

Lawler, J. J., White, D., Neilson, R. P., & Blaustein, A. R. (2006). Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biology*, *12*(8), 1568-1584.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

Li, Y., & Lindstrom, M. J. (2001). Evaluating soil quality–soil redistribution relationship on terraces and steep hillslope. *Soil Science Society of America Journal*, *65*(5), 1500-1508.

Liu, J., Goering, C. E., & Tian, L. (2001). A neural network for setting target corn yields. *Transactions of the ASAE*, *44*(3), 705.

Lobell, D. B., & Field, C. B. (2011). California perennial crops in a changing climate. *Climatic Change*, *109*(1), 317-333.

Longman, R.H.G., Ter Braak, C.J.F., Van Tongeren, O.F.R. (1995). Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge.

Lowenberg-DeBoer, J., & Erickson, B. (2019). Setting the record straight on precision agriculture adoption. *Agronomy Journal*, *111*(4), 1552-1569.

Mahler, R. L., Bezdicek, D. F., & Witters, R. E. (1979). Influence of Slope Position on Nitrogen Fixation and Yield of Dry Peas 1. *Agronomy Journal*, *71*(2), 348-351.

Mallarino, A. P., Beegle, D. B., & Joern, B. C. (2007). Soil Sampling Methods for Phosphorus-Spatial Concerns. A SERA-17 Position Paper.

Marinković, B., Crnobarac, J., Brdar, S., Antić, B., Jaćimović, G., & Crnojević, V. (2009, October). Data mining approach for predictive modeling of agricultural yield data. In *Proc. First Int Workshop on Sensing Technologies in Agriculture, Forestry and Environment (BioSense09), Novi Sad, Serbia* (pp. 1-5).

McBratney, A., Whelan, B., Ancev, T., & Bouma, J. (2005). Future directions of precision agriculture. *Precision agriculture, 6*(1), 7-23.

McConkey, B. G., Ulrich, D. J., & Dyck, F. B. (1997). Slope position and subsoiling effects on soil water and spring wheat yield. *Canadian journal of soil science, 77*(1), 83-90.

McKenzie, N., Jacquier, D., Isbell, R., & Brown, K. (2004). *Australian soils and landscapes: an illustrated compendium*. CSIRO publishing.

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., & Van Molle, M. (2008). A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma, 143*(1-2), 1-13.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics, 10*(1), 213.

Metwally, M. S., Shaddad, S. M., Liu, M., Yao, R. J., Abdo, A. I., Li, P., … Chen, X. (2019). Soil properties spatial variability and delineation of site-specific management zones based on soil fertility using fuzzy clustering in a hilly field in Jianyang, Sichuan, China. *Sustainability (Switzerland)*, *11*(24).

Miao, Y., Mulla, D. J., & Robert, P. C. (2006). Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. *Precision Agriculture, 7*(2), 117–135.

Mishra, A. K., Nimon, R. W., & El-Osta, H. S. (2005). Is moral hazard good for the environment? Revenue insurance and chemical input use. *Journal of environmental management*, *74*(1), 11-20.

Mittal, G. S., & Zhang, J. (2000). Prediction of temperature and moisture content of frankfurters during thermal processing using neural network. *Meat Science*, *55*(1), 13-24.

Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological processes*, *5*(1), 3-30.

Moore, I. D., Gessler, P. E., Nielsen, G. A. E., & Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil science society of america journal*, *57*(2), 443-452.

Moore, G. A. (2001). *Soilguide (Soil guide): A handbook for understanding and managing agricultural soils*.

Mousavi, S. R., Galavi, M., & Rezaei, M. (2013). Zinc (Zn) importance for crop production—a review. *International Journal of Agronomy and Plant Production*, *4*(1), 64-68.

Mzuku, M., Khosla, R., Reich, R., Inman, D., Smith, F., & MacDonald, L. (2005). Spatial Variability of Measured Soil Properties across Site-Specific Management Zones. *Soil Science Society of America Journal*, *69*(5), 1572–1579.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

Odeha, I. O. A., McBratney, A. B., & Chittleborough, D. J. (1994). Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, *63*(3-4), 197-214.

[OMAFRA] Ontario Ministry of Agriculture, Food and Rural Affairs. (2009). Agronomy Guide for Field Crops. Publication 811.

Ontario Ministry of Natural Resources. (2015). Southwestern Ontario Orthophotography Project (SWOOP) - Imagery Package D.

Ovalles, F. A., & Collins, M. E. (1986). Soil-landscape relationships and soil variability in north central Florida. *Soil Science Society of America Journal*, *50*(2), 401-408.

Panagopoulos, T., Jesus, J., Antunes, M. D. C., & Beltrao, J. (2006). Analysis of spatial interpolation for optimising management of a salinized field cultivated with lettuce. *European Journal of Agronomy*, *24*(1), 1-10.

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, *121*, 57-65.

Pasolli, L., Notarnicola, C., & Bruzzone, L. (2011). Estimating soil moisture with the support vector regression technique. *IEEE Geoscience and remote sensing letters*, *8*(6), 1080-1084.

Pathak, H. S., Brown, P., & Best, T. (2019). A systematic literature review of the factors affecting the precision agriculture adoption process. *Precision Agriculture*, *20*(6), 1292-1316.

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Paxton, K. W., Mishra, A. K., Chintawar, S., Larson, J. A., Roberts, R. K., English, B. C., ... & Martin, S. W. (2010). *Precision agriculture technology adoption for cotton production* (No. 1370-2016-108739).

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Prediction and explanation. *New York: Holt, Rinehart, & Winston*.

Pennock, D. J., & Jong, E. D. (1990). Rates of soil redistribution associated with soil zones and slope classes in southern Saskatchewan. *Canadian Journal of Soil Science*, *70*(3), 325-334.

Pettigrew, W. T. (2008). Potassium influences on yield and quality production for maize, wheat, soybean and cotton. *Physiologia plantarum*, *133*(4), 670-681.

Pierce, F. J., & Nowak, P. (1999). Aspects of precision agriculture. In *Advances in agronomy* (Vol. 67, pp. 1-85). Academic Press.

Rainbow, R., Wells, M. (2004). Precision agriculture and remote sensing workshops. October, Southern Precision Agriculture Association, Clare, South Australia.

Raorane, A. A., & Kulkarni, R. V. (2012). Data Mining: An effective tool for yield estimation in the agricultural sector. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, *1*(2), 1-4.

Rayment, G. E., & Higginson, F. R. (1992). *Australian laboratory handbook of soil and water chemical methods*. Inkata Press Pty Ltd.

Richards, N. R., Caldwell, A. G., & Morwick, F. F. (1949). *Soil survey of Essex County* (No. 11). Experimental Farms Service, Dominion Department of Agriculture and the Ontario Agricultural College.

Ristow, P. L., Foster, J., & Ketterings, Q. M. (2010). Lime guidelines for field crops; tutorial workbook. *Department of animal science. Cornell University, Ithaca*, *47*.

Robertson, M., Isbister, B., Maling, I., Oliver, Y., Wong, M., Adams, M., et al. (2007). Opportunities and constraints for managing within-field spatial variability in Western Australian grain production. Field Crops Research, 104(1–3), 60–67.

Robson, A. D., & Pitman, M. G. (1983). Interactions between nutrients in higher plants. In *Inorganic plant nutrition* (pp. 147-180). Springer, Berlin, Heidelberg.

Ruß, G. (2009, July). Data mining of agricultural yield data: A comparison of regression models. In *Industrial Conference on Data Mining* (pp. 24-37). Springer, Berlin, Heidelberg.

Ruß, G., & Kruse, R. (2010). Feature selection for wheat yield prediction. In *Research and Development in Intelligent Systems XXVI* (pp. 465-478). Springer, London.

Sadler, E. J., Sudduth, K. A., & Jones, J. W. (2007). Separating spatial and temporal sources of variation for model testing in precision agriculture. *Precision Agriculture*, *8*(6), 297-310.

Safa, B., Khalili, A., Teshnehlab, M., & Liaghat, A. (2004). Artificial neural networks application to predict wheat yield using climatic data. In *Proceedings of 20th International Conference on IIPS* (pp. 1-39). Iranian Meteorological Organization.

Sandri, M., & Zuccolotto, P. (2010). Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms. *Statistics and Computing*, *20*(4), 393-407.

Sarvari, H., & Keikha, M. M. (2010, December). Improving the accuracy of intrusion detection systems by using the combination of machine learning approaches. In *2010 international conference of soft computing and pattern recognition* (pp. 334-337). IEEE.

Schieffer, J., & Dillon, C. (2013). Precision agriculture and agro-environmental policy. In *Precision agriculture'13* (pp. 755-760). Wageningen Academic Publishers, Wageningen.

Schlenker, W., & Roberts, M. J. (2006). Nonlinear effects of weather on corn yields. *Review of agricultural economics*, *28*(3), 391-398.

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of the National Academy of sciences*, *106*(37), 15594-15598.

Seyhan, A. T., Tayfur, G., Karakurt, M., & Tanogˇlu, M. (2005). Artificial neural network (ANN) prediction of compressive strength of VARTM processed polymer composites. *Computational Materials Science*, *34*(1), 99-105.

Shearer, S. (2014). The Ag Info Conference 2014. Retrieved September 24, 2015, from Big Data: The Future of Precision Agriculture http://infoag.org/abstract_papers/papers/paper_233.pdf

Shearer, S. A., Burks, T. F., Fulton, J. P., Higgins, S. F., Thomasson, J. A., Mueller, T. G., & Samson, S. (1999). Yield prediction using a neural network classifier trained using soil landscape features and soil fertility data. In *ASAE Paper* (pp. 99-3042).

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M., & Ebrahimie, E. (2014). Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture. *PloS one*, *9*(5).

South, D. B., & Davey, C. B. (1983). Southern forest nursery soil testing program.

Stubbs, M. (2016). *Irrigation in US agriculture: on-farm technologies and best management practices*. Washington, DC: Congressional Research Service.

Sudduth, K. A., Drummond, S. T., Birrell, S. J., & Kitchen, N. R. (1996). Analysis of spatial factors influencing crop yield. *Precision Agriculture*, (precisionagricu3), 129-139.

Subhadra, M., Debahuti, M., Gour Hari, S. (2016). Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper. Indian J. Sci. Technol. 9.

Sumner, M. E., & Farina, M. P. (1986). Phosphorus interactions with other nutrients and lime in field cropping systems. In *Advances in soil science* (pp. 201-236). Springer, New York, NY.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.

Svetnik, V., Liaw, A., & Tong, C. (2004). Variable selection in random forest with application to quantitative structure-activity relationship. *Proceedings of the 7th Course on Ensemble Methods for Learning Machines*.

Tan, C. S., & Reynolds, W. D. (2003). Impacts of recent climate trends on agriculture in southwestern Ontario. *Canadian Water Resources Journal*, *28*(1), 87-97.

Tantalaki, N., Souravlas, S., & Roumeliotis, M. (2019). Data-Driven Decision Making in Precision Agriculture: The Rise of Big Data in Agricultural Systems. *Journal of Agricultural & Food Information*, *20*(4), 344-380.

Tey, Y. S., & Brindal, M. (2012). Factors influencing the adoption of precision agricultural technologies: A review for policy implications. *Precision Agriculture*, *13*(6), 713–730.

Understanding a semivariogram: The range, sill, and nugget. (2020). Retrieved from https://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/understanding-a-semivariogram-the-range-sill-and-nugget.html

Urban, D., Roberts, M. J., Schlenker, W., & Lobell, D. B. (2012). Projected temperature changes indicate significant increase in interannual variability of US maize yields. *Climatic change*, *112*(2), 525-533.

Utset, A., Ruiz, M. E., Herrera, J., & de Leon, D. P. (1998). A geostatistical method for soil salinity sample site spacing. *Geoderma*, *86*(1-2), 143-151.

Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J. P., Veroustraete, F., Clevers, J. G., & Moreno, J. (2015). Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties–A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *108*, 273-290.

Wang, Y., Witten, I. H., van Someren, M., & Widmer, G. (1997). Inducing models trees for continuous classes. In *Proceedings of the Poster Papers of the European Conference on Machine Learning, Department of Computer Science, University of Waikato, New Zeland*.

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for

predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators*, *52*, 394-403.

Wilson, J. P., & Gallant, J. C. (2000). Digital terrain analysis. *Terrain analysis: Principles and applications*, *6*(12), 1-27.

Wollenhaupt, N. C., & Wolkowski, R. P. (1994). Grid soil sampling. *Better Crops*, *78*(4), 6-9.

Wright, R. J., Boyer, D. G., Winant, W. M., & Perry, H. D. (1990). The influence of soil factors on yield differences among landscape positions in an Appalachian cornfield. *Soil Science*, *149*(6), 375-382.

Veenadhari, S., Misra, B., & Singh, C. D. (2011). Data mining techniques for predicting crop productivity—A review article. *IJCST*, *2*(1), 90-100.

Verity, G. E., & Anderson, D. W. (1990). Soil erosion effects on soil quality and yield. *Canadian Journal of Soil Science*, *70*(3), 471-484.

Yadav, R., Rathod, J., & Nair, V. (2015). Big data meets small sensors in precision agriculture. *International Journal of Computer Applications*, *975*, 8887.

Yang, C., Peterson, C. L., Shropshire, G. J., & Otawa, T. (1998). Spatial variability of field topography andwheat yield in the palouse region of the Pacific Northwest. *Transactions of the ASAE*, *41*(1), 17.

Yosefi, K., Galavi, M., Ramrodi, M., & Mousavi, S. R. (2011). Effect of bio-phosphate and chemical phosphorus fertilizer accompanied with micronutrient foliar

application on growth, yield and yield components of maize (Single Cross 704). *Australian journal of crop science*, *5*(2), 175.

Yost, M. A., Russelle, M. P., Coulter, J. A., Sheaffer, C. C., & Kaiser, D. E. (2011). Potassium management during the rotation from alfalfa to corn. *Agronomy journal*, *103*(6), 1785-1793.

Zhang, L., Zhang, J., Kyei-Boahen, S., & Zhang, M. (2010). Simulation and prediction of soybean growth and development under field conditions. *Am-Euras J Agr Environ Sci*, *7*(4), 374-385.

Zhang, N., Wang, M., & Wang, N. (2002). Precision agriculture—a worldwide overview. *Computers and electronics in agriculture*, *36*(2-3), 113-132.