Applying machine learning digital soil mapping techniques on farm scale for use in precision agriculture

EH Louw

orcid.org 0009-0001-0280-2643

Dissertation accepted in partial fulfilment of the requirements for the degree *Master of Science in Agriculture with Soil Science* at the North-West University

Supervisor: Prof GM van Zijl

Graduation: July 2025

ACKNOWLEDGEMENTS

Firstly, I would like to give thanks to the one true God: Father, Son and Holy Spirit, for his guidance and support on my journey through my Magister Scientia in Agriculture. It is through Him and Him alone that I had the privileged life filled with opportunity that I have lived. I hope that the effort and thought put into this dissertation bring glory to His name.

Psalm 127 1-2 (AFR53): "*As die Here die huis nie bou nie, tevergeefs werk die wat daaraan bou; as die Here die stad nie bewaar nie, tevergeefs waak die wagter. Tevergeefs dat julle vroeg opstaan, laat opbly, brood van smarte eet — net so goed gee Hy dit aan sy beminde in die slaap!"*

Thank you to the people and organisations who were instrumental to this research project:

- I would like to thank my supervisor, Prof. GM van Zijl, for his guidance and support during the writing of this dissertation. The support and feedback you gave me during our consultations were instrumental to my personal and professional life during this time. I will forever be grateful for your support and kindness during this interesting and sometimes hard time.
- I would like to thank my parents, Eric and Driekie Louw, for their support and encouragement throughout this endeavour. Also, my sister, Marlizé Louw, who supported and encouraged me whilst also working on her own dissertation.
- NWK for providing the soil chemical analyses and soil physical data.
- Dieter Hansen for providing access to his fields and yield data.
- Willie Cloete and Anru Kock, PhD students, for assisting me with programming, researching and writing.

ABSTRACT

This study aimed to evaluate the industry-standard method of ordinary kriging (OK) used for creating soil chemical property maps and then also to investigate the potential of using machine learning-based digital soil mapping techniques to create the same maps for use in precision agriculture (PA). To do this, two fields in the Gerdau area in the North West of South Africa were selected, and soil chemical analyses for them on a 2-ha grid were obtained from NWK. Using similar processes to the industry standard, OK maps for pH, P, K, Ca, Mg and Na were produced, and the accuracy thereof was tested using leave-one-out cross-validation. The semivariograms of each property were also investigated to determine the optimal sampling density. After this, the same sets of maps were created using the ML DSM techniques Cubist and Random Forest (RF), the accuracy of which was tested by splitting the data into a training and validation set. The ML DSM maps were then also compared to the OK maps through visual inspection. It was found that soil properties varied between Fields and each other. In Field 1, only the OK map of Mg was somewhat accurate enough for use in PA. More success was found in Field 2 with the OK maps of pH, K and Mg being accurate enough for use in PA. Based on the semivariograms of these properties, a smaller inter-sample range is needed than the 2-ha grid provides. Based on the results of this study, it is recommended that sample grids no larger than 70 by 70 meters be used, in line with the findings of Brouder and Morgan (2000). None of the ML DSM maps proved more accurate than the OK maps on field scale in this case, meaning that legacy soil and yield data could not be used on field scale to create comparable soil property maps for use on field scale PA.

Key terms: Digital soil mapping, Ordinary Kriging, Soil Mapping, Precision Agriculture, Machine learning, Soil properties

LIST OF ABBREVIATIONS

Ca	Calcium
CCC	Lin’s concordance correlation coefficient
CEC	Cation exchange capacity
cLHS	Conditioned Latin hypercube sampling.
Cv	Coefficient of variability
DSM	Digital soil mapping
FKM	Fuzzy K-means clustering
GIS	Geographic information system
GPS	Global Positioning System
K	Potassium
KCl	Potassium chloride
KM	K-means clustering
MAE	Mean absolute error.
ML	Machine learning
MLR	Multinomial logistic regression
MLRM	Multilinear regression models
Mg	Magnesium
MIR	Mid Infrared
LULC	land use/ land cover
Na	Sodium
OK	Ordinary Kriging
P	Phosphorus
PA	Precision Agriculture
RF	Random Forest
RK	Regression Kriging
RMSE	Root mean square error.

RS	Remote sensing
RTs	Regression trees
SOC	Soil organic carbon
SOM	Soil organic matter
SSCM	Site specific crop management
TN	Total Nitrogen

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Introduction

Agriculture plays a very important role in South Africa's economy and food systems. As more yields become necessary, several food production systems have become popular, one of which is Precision Agriculture (PA), which makes use of timely soil and environmental data to determine production strategies. Using these data sets, maps are created to guide management decisions and strategies for the site-specific management (SSCM) of crops. It is thus important that these maps are created promptly and with enough accuracy. As such, new and novel methods of soil property mapping are constantly being explored, and existing methods are being tested. In South Africa, soil chemical property mapping is usually done with ordinary kriging (OK) and regression kriging (RK), but digital soil mapping (DSM) methods have become a popular alternative for researchers (Nenkam *et al.,* 2024).

1.2 Background

As the world population continues to increase, more food needs to be produced. As such, new and innovative sustainable food production methods are being developed and refined. One of these methods is PA. PA is a method of crop production based on the minimisation of inputs while maximising yield through the management of variability within the field (Linden, 2003). The concept of PA is also more broadly defined by Whelan and Taylor (2013), as a form of whole farm management strategies that focusses on achieving production efficiency by minimising wasteful inputs and unintended environmental impacts by making use of information technology. In this case SSCM becomes a form of PA in crop production systems where infield variability is measured and management practices are adapted to better address this variability (Whealan and Taylor, 2013). As this study is concerned with PA in the grain crop production systems of South Africa, the term's use is solely in conjuncture with the concepts of SSCM. The variability in fields can be managed through two possible methods, as described by Zhang et al. (2002), either a map- or sensor-based approach. The map-based approach is the more popular method, thanks to readily available technologies like Global Positioning Systems (GPS), yield monitoring, satellite imaging, and soil sampling methods. The creation of maps used in this approach needs to be made accurately and timely manner to be of use to PA's decision-making support process (Bobryk et al., 2016). Soil sampling is usually done in a grid-based system with one or two-hectare grid spacing, though it has been realised that this sampling density can be too low because of the high variability of some soil properties (Linden, 2003). This sampling method can also be costly and time-consuming (Van Zijl et al., 2013). A possible solution to this problem is the implementation

of ML based digital soil mapping (DSM) techniques to create soil maps using legacy data and thus reducing the need for extensive soil sampling. As DSM projects have usually been done on a regional scale (McBratney et al., 2003), this study focuses on creating soil property maps suitable for PA, using fewer observations, and using ML-based DSM on field scale.

1.3 Problem statement

The ever-increasing population of sub-Saharan Africa and the world necessitates more and more food admits ever-increasing input costs. Sustainability has thus become more important than ever. PA has the potential to serve as a sustainable method of agricultural production (Zhang et al., 2002; Bongiovanni & Lowenberg-DeBoer, 2004; Van Evert et al., 2017 & Loures et al., 2020). This is mainly done through the minimisation of inputs and maximisation of outputs through decision-making based on spatial and temporal information to manage variability within crop production systems (Gebbers & Adamchuk, 2010). A central pillar of this decision-making system is spatial soil maps, which greatly dictate the distribution of management zones (Bobryk et al., 2016). These maps need to be created accurately and timely manner to be useful (van Zijl, 2019). One of the problems faced by PA is that conventional grid-based surveys are costly and time-consuming (Van Zijl et al., 2013). In South Africa, these surveys are usually done in one- or two-hectare grids that then get interpolated (van Zijl, 2019). In the context of this study the industry method being tested is OK. During the spatial interpolation process, discrete point-based data gets transformed into continuous soil property maps (Schloeder et al., as referenced by ZHU and Lin (2010)). This is made possible by the fact that at any spatial and temporal point, soil attributes can be empirically described as a function of the surrounding environment (Boettinger et al., 2008). Unfortunately, the grid resolution lacks the data to accurately predict all soil characteristics, as the interpolation process is very sensitive to the number of samples and the spatial variation thereof (ZHU & Lin, 2010). A possible solution to this problem is making use of machine learning based digital soil mapping (MLDSM) techniques to better predict soil characteristics based on environmental factors available at the farm scale. By increasing the accuracy of soil maps and reducing the time needed to create a map, agricultural producers should be able to make better and more timely decisions. Generally, DSM projects have focused on a regional scale (McBratney et al., 2003). This study is aimed at creating soil property maps suitable for PA, using fewer soil observations, due to utilising the methods of ML DSM. So that the accuracy of the industry standard method's accuracy can be tested and compared to the possible alternative to interpolation modelling of ML DSM.

1.4 Hypothesis, research aims and objectives

1.4.1 Hypothesis

The soil spatial maps created through ML DSM techniques will be equally accurate to the current South African industry standard, whilst also being more applicable.

1.4.2 Research aims

The study aims to evaluate the industry standard method used for creating soil property maps, such as used by a leading company providing fertiliser recommendations, namely OK, and to test the potential of using ML DSM with legacy data to create the same maps for use in PA.

1.4.3 Research objectives

The objectives that need to be met to achieve the aim of the study are:

1) Test whether the conventional, industry-standard OK method on a 2-ha grid is sufficient for use in PA.

2) Determine the optimal grid spacing for soil property mapping using OK for PA.

3) Determine the potential for use of ML DSM for mapping soil properties in PA.

By making use of legacy data provided by a leading company providing fertiliser recommendations.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

During this study, the focus is on the use of ML methods, such as Cubist and random forest (RF), and traditional interpolation techniques like Ordinary Kriging (OK) in mapping of soil chemical properties using DSM for use in PA. The literature review includes descriptions, developments and evaluations of these techniques in PA in the context of central South Africa at field scale.

2.2 Precision agriculture

2.2.1 What is PA?

PA is a term that can be used to describe several production strategies within the wide field of agriculture. The use of the term is usually understood in the context of production system being spoken of although

PA is a system of food production that makes use of technology to capture the in-field variability of yield-influencing environmental factors to then create models thereof using statistical methods to be used for the creation of decision-making support systems (Linden, 2003). These systems are then used to implement site-specific management on crop production systems in a timely manner (Gebbers & Adamchuk, 2010). Thus, by adapting inputs to site-specific in-field applications, production can be made more efficient by reducing wasteful fertiliser application (Van Evert et al., 2017). Thus, profit is increased while environmental impact is kept to a minimum. As such PA can be defined as a crop production method that makes use of auxiliary variables, measured by advanced technologies, and expert knowledge systems to make timely decisions based on the measurable-, natural- and anthropogenic- environmental variability to adjust site specific production methods as to optimize the profitability, production quantity and environmental impact of the entire crop production process.

The order of operation for PA is thus the collection of site-specific crop and soil data, the presentation and interpretation of the data, decision-making, modelling site-specific operations, and finally the application of a site-specific treatment. The most important information technologies used in PA are GPS, multimedia, remote sensors, unmanned aerial and land vehicles, variable rate technology, and wireless sensors (Cisternas et al., 2020).

Zhang et al. (2002) identify six main variable groups that significantly influence agricultural production: Yield-, field-, crop-, anomalous factors- and management-variability. Yield variability includes past and present yield distributions; field variability includes topography, elevation, slope, aspect, and terrace; soil variability can include soil nutrient content, moisture content, other physical properties like texture and chemical properties like pH; crop variability includes plant-height, density, nutrient- and water- stress levels and other biophysical properties; anomalous factors include the various environmental disasters and pest infestations the crop can be subject to; and lastly management variability is characterized as the tillage- and planting- practices applied as well as the applications of herbicides, pesticides, and fertilizers (Zhang et al., 2002).

In order to manage variability, a map-based approach can be taken, wherein fields are sampled in a pre-decided manner, the samples are analysed, site-specific maps are created, and treatment is applied in a site-specific manner (Zhang et al., 2002). Soil samples were noted as an important factor for comparison, by Burlacu *et al.* (2014)), because of its ability to give a general description of soil physical and chemical property variance.

2.2.2 PA in the South African context

As mentioned earlier, PA can have a wide range of definitions, but in the context of grain production in South Africa, it is usually used to describe a method of SCCM. This is evident by the use of terms like “Precision Service Mapping” by the fertiliser company Kynoch (personal correspondence) and “Precision Analyses” by NWK (personal correspondence). This Jacobs et al. (2018) found that the main use of PA in the Schweizer-Reneke area, within 130 kilometres of the study site of this thesis, in South Africa, is for fertiliser recommendation and that the dominant factor controlling the adoption of PA is farm size. All farmers in the study area adopted PA if their crop land exceeded 3,000 ha, while farmers with less than 1500 ha did not adopt PA at the time of the study. In the same study, Jacobs et al. (2018) mention that the main advantage of PA experienced by farmers was the increase in potential yield and realised yield within two years of adoption. According to a study done by Nyaga et al. (2021) on PA research in Sub-Saharan Africa, the main focus of PA research in the area is soil mapping, with the majority of studies on geographic information systems (GIS) and remote sensing (RS) done on large-scale farms.

2.3 Conventional soil mapping in South Africa

In the case of soil variability, samples are usually taken as points based on a grid of one hectare, in Nordic countries (Linden, 2003), and one or two hectares in South Africa (van Zijl, 2019). Brouder & Morgan (2000) recommend that grids no larger than 70 meters be used in the United States of America. As for the depth of sampling, it is recommended that topsoil samples be taken from 0 to 30 cm depth (Fertiliser Society of South Africa, 2016). These varying sampling densities exist because soil properties can vary significantly over relatively short distances within the same field (Paterson et al., 2015). This variability can range from micro- to mega-scopic in scale (Keskin & Grunwald, 2018). In South Africa, significant variation of pH, K, P, Ca and Mg levels between crop rows and inter-crop rows was found by Van Vuuren et al. (2000). This research formed the backbone of NWK’s soil sampling method, which imitates the trench method. In this method, five samples are taken across the crop row diagonally, from 0 to 30 cm depth, and a single composite sample of around 500 g is taken for analysis according to du Plessis, MJ. (2018). Du Plessis (2018) further states that this is done as it is in this zone that most of the crop’s nutrients are to

be acquired. NWK does their soil analyses for PA on grid patterns with sampling densities of one and two hectares (NWK, 2024). Soil properties are then mapped using OK according to personal correspondence with NWK. Although more sampling methods have been advocated for, no evidence has been presented for widespread use thereof within South Africa's agriculture sector. Some of the methods of include a stratified random point survey, an area composite method wherein fields are subdivided into management zones, and field zone methods based on ancillary data (Linden, 2003).

Unfortunately, traditional soil sampling methods are costly and time-consuming (Van Zijl et al., 2013). This is because developing the skills and knowledge needed to accurately collect soil information is labour-intensive and time-consuming, as is the data collection process itself (Paterson et al., 2015). There is a valid further concern that the industry standard of one- or two-hectare grids is not a sufficiently dense sampling design to accurately describe soil property variation (Linden, 2003). Linden (2003) also found that increasing the sampling density significantly increased accuracy.

2.4 Soil property mapping methods

2.4.1 Introduction

After sampling and analysis, maps are usually generated using DSM or geostatistical techniques. Lagacherie et al. (2006), as referenced by Pereira et al. (2022), described DSM as the process wherein spatial soil information systems as derived from field observations and laboratory analyses combined with inference systems. This is possible because soil properties and their distribution in a landscape are dependent on the environmental factors under which they are formed (Paterson et al., 2015). These environmental factors mainly consist of the five major soil-forming factors (climate, organisms, parent material, relief, parent material and time) as described by McBratney et al. (2003), referenced by Söderström et al. (2016). Taken further the soil class/property to be mapped (S) is a function of the soil (s), climate (c), organic matter (o), relief (r), parent material (p), age (a) and spatial variability (n), (McBratney *et al.*, 2003), as displayed in Equation 2.4.1.

Equation 2.4.1: Soil as a function of its forming factors

$$S = f(s, c, o, r, p, a, n)$$

To create a DSM, enough soil and environmental data are needed to describe the soil-environmental relationships (Paterson et al., 2015). At that point, it becomes possible to use the

environmental covariates to predict soil distribution at unobserved locations (Van Zijl et al., 2020). It also allows for the creation of continuous soil property maps based on discrete data (Paterson et al., 2015). The two options for modelling soil variation are thus discrete modelling (polygon-based) and continuous modelling (pixel-based) (Keskin & Grunwald, 2018). In the polygonal approach, soils are partitioned into homogeneous classes and in the pixel-based approach, they are elucidated into a continuum (Keskin & Grunwald, 2018). Three DSM methods exist, namely geostatistical methods, expert knowledge base and ML (Bobryk et al., 2016).

2.4.2 Kriging

2.4.2.1 Ordinary Kriging

Geostatistical methods used commonly in agriculture include OK and RK (ZHU & Lin, 2010). These methods make use of interpolation models to map soil distribution patterns over multiple scales of time and space (Keskin & Grunwald, 2018). This is possible because the values of soil properties observed at close geographical spacing are similar or spatially correlated, as observed by Oliver (1987)The prediction of soil chemical properties is done through a weighted local averaging that makes use of a variogram, representing the measure of spatial dependence, to assign importance to data points when calculating the averages in between (Scull et al., 2003). These methods are very spatially a quantitatively dependent (ZHU & Lin, 2010), meaning that the distance between observations, their location in the space they exist in and the number of observations within the location greatly influence the performance of these models.

A common criticism of geostatistical methods, especially of methods like OK and RK is that being a global instead of a local technique, it fails to take knowledge of soil materials and processes into consideration (Scull et al., 2003). This then means that when the values of single pixels are calculated, the global trend has more weight in the calculation than the environmentally related variance in the soil. Two of the most common geostatistical methods used in agriculture are OK and RK (ZHU & Lin, 2010). The latter makes use of auxiliary variables to improve interpolation accuracy (ZHU & Lin, 2010). However, this requires even more data, where initial data points are already scarce. Co-kriging is another kriging method, though less commonly used, that makes use of auxiliary variables, like remote sensing data, to make more accurate predictions (Knotters et al., 1995)

2.4.2.2 The Semi variogram

The semi-variogram (Figure 2.1) is used to characterise certain soil attributes as a function of the distance between adjacent, stationary points (Keskin & Grunwald, 2018). The experimental

semivariogram can be calculated with the following Equation 2.4.2, as presented by Suleymanov et al. (2023).

Equation 2.4.2: The experimental semivariogram

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

γ(h) represents semi-variance at a distance of h. The number of pairs of sample points at a distance of h is given as N(h). Z(xi) is the value of the variable (Z) at a set location (xi). The variance is then determined for each possible pair within a pre-determined distance, chosen by the user. In this way, a global model representing all points within a field can be used, or a local model based on only observing points in a set search distance (Whelan & Taylor, 2013). The variogram model describes how the variation $\gamma(h)$ between data points changes as the distance between points (N(h)) increases. Figure 2.1, based on the descriptions of Whelan and Taylor (2013) and Lamine *et al.* (2023), represents an almost perfect Semivariogram, the exact shape of which changes with different model types.

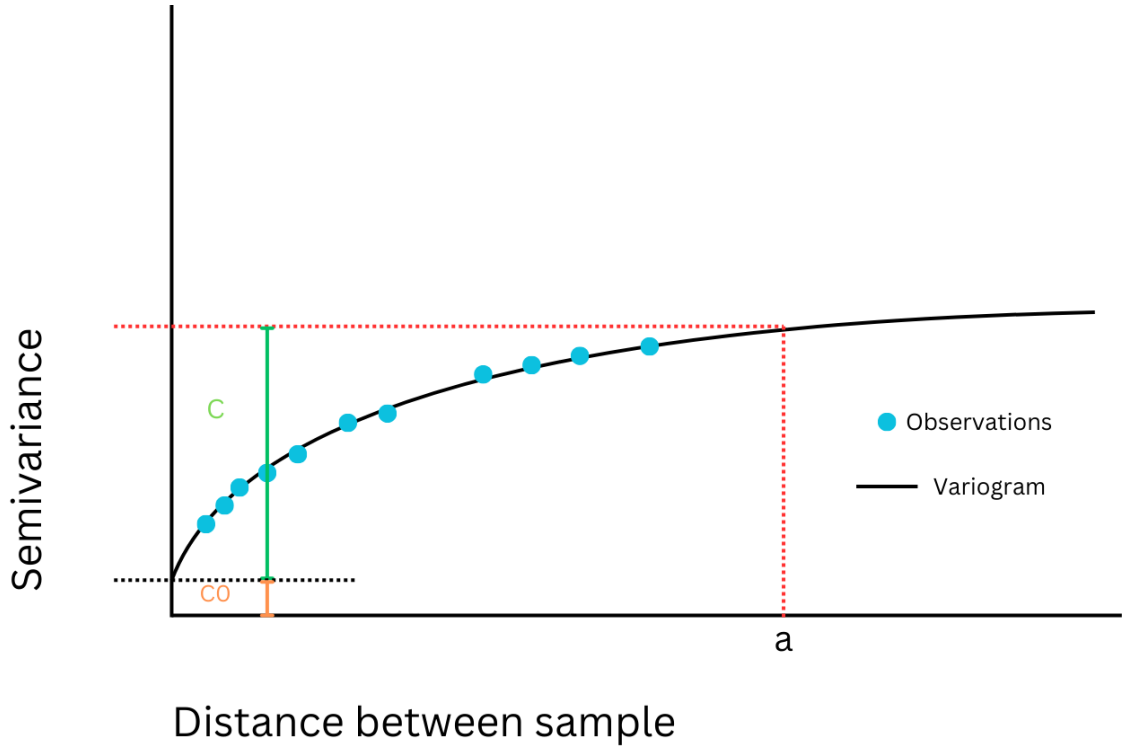

Figure 2.1: Representation of the Semivariogram model based on descriptions from Whelan and Taylor (2013) and Lamine *et al.* (2023). C0 = Nugget variation (close variation); C = Spatialy dependant variation; C0+C = maximum variability; a = range of spatial dependence

In Figure 2.1, any points within the range (a) from each other can still be predicted with a good measure of accuracy by the semivariogram model, if the value of any of the other points in that

range is known. As the distance between points increases, less weight is given to the probability that they are the same. If the semivariogram model accurately describes the distance-dependent spatial variation between observations, there will be minimal differences between the variogram and the observations, but if no distance-dependent spatial variation can be found, the two will differ in shape or trend. This quantification of the relation between distance and similarity is commonly referred to as spatial autocorrelation and is foundational to kriging (Tobler, 1970). To determine the required sampling density, a variogram analysis can be done, although 1.5 to 2 ha per observation point is generally considered to be sufficient (Linden, 2003). Commonly referred to as uncertainty estimation, this involves investigating how model parameters and sampling designs effect the prediction reliability of a variogram by looking at how these factors influence variance (Goovaerts, 1997). The ability of a semivariogram to explain variance is also directly dependant on the quality of the data. High quality spatial data needs to be accurate, consistent and complete which can be assessed by looking for measurement errors, assessing spatial coverage and evaluating sampling density Goodchild & Li (2012). Bad spatial data quality would make it hard to determine the effectiveness of any model.

2.4.3 Machine learning DSM

The ML approach reduces the human input by automating the process of linking soil properties and environmental factors through algorithm building done by a computer based on large data sets (Pereira et al., 2022). The creation of the rule is thus handled by the machine instead of the expert. This is done by making use of data mining techniques to identify patterns (learning) and then building models (Khaledian & Miller, 2020). These non-linear statistical methods have shown that they could be better at explaining complex soil-environment relationships than linear statistical methods (Huang et al., 2022). This is likely because of the ability of ML models to handle non-normally distributed and hierarchically related data (Keskin & Grunwald, 2018). Two ML methods that are good at handling smaller data sets are Random Forest (RF) and Cubist, especially for less than 100 data points (Khaledian & Miller, 2020). Both methods make use of decision trees when making predictions.

2.4.3.1 Cubist

According to Khaledian and Miller (2020), Cubist uses the provided covariates to create an initial tree structure; the branches of this tree then get collapsed as rules are created by the algorithm through boosted training. Boosting is "*an approach that converts weak learners to strong learners by applying more weight on the stronger learners*", according to Khaledian and Miller (2020). They

further explain that each rule is a multinomial logistic regression (MLR) model, which uses the conditions under which that rule is formed to predict the target variable. Starting from the top of the tree, each node smooths the prediction to decrease the prediction error based on the prediction from previous nodes (Khaledian & Miller, 2020). At the end of the process, a final prediction model is generated consisting of a collection of multiple linear regression models (MLRM) (Khaledian & Miller, 2020). It is this collection of MLRMs, used to create a single predictive model, that makes Cubist and RF an ensemble learning method (Khaledian & Miller, 2020). Khaledian and Miller (2020) found that the Cubist algorithm has been largely used internationally for the prediction of the distribution of soil organic carbon (SOC), bulk density, soil particle size distribution, and horizon thickness. A recent example is that of the study by Suleymanov *et al.* (2023) where the Cubist model, amongst others, was used to create prediction maps for the distribution of soil organic matter (SOM) percentage, pH (H2O) and pH (KCl) on a 5 Km² area using DEM, Landsat 2 spectral data and land use/ land cover (LULC) raster data. They found that SOM could be predicted and mapped with an accuracy of R^2 of 0.64 and RMSE of 1.95, while pH (KCl) R^2 of 0.34 and RMSE of 0.66. Within South Africa, Cubist has been used on field scale with mid-infrared (MIR) spectroscopy models, by Kock *et al.* (2024) to predict soil chemical properties such as pH, P, K, Ca, Mg and Na in the western highveld region. The most accurate prediction map they produced with Cubist was for pH with an R^2 of 0.86 and RMSE of 0.29, followed by Ca with an R^2 of 0.84 and RMSE of 90.32.

2.4.3.2 Random Forest

According to Taghizadeh Mehrjardi et al (2016), Pahlavan-Rad et al. (2018) and Huang et al. (2022), RF is the combination of regression trees (RTs) that then give a single prediction. These RTs are generated by making use of randomly selecting observations (Khaledian & Miller, 2020) and then aggregating them (Dong et al., 2019). Two-thirds of the observations are used with a randomised arrangement of covariate orders, whilst the rest is left out for determining model error, known as an out-of-bag strategy (Khaledian & Miller, 2020). By making use of bootstrap aggregating, RF reduces variance and improves the stability of prediction results (Khaledian & Miller, 2020). The final predictions are then made based on the mean of results produced by all the decision trees (Khaledian & Miller, 2020).

In comparison to Cubist, RF is a more popular ML DSM method. On the regional scale, it has been used by Xiao *et al.* (2023) to map soil pH (H_2O) levels across Europe with an R^2 of 0.68 and RMSE of 0.76. In China Dong *et al.* (2019) used RF to map SOM, total nitrogen (TN), available P and available K on a 30 m grid over a large area of the Zhongning county using soil and environmental covariates like, amongst others, soil type and elevation. They reported the highest

R^2 for the TN map at 0.43 with an RMSE of 4.23, followed by SOM with an R^2 of 0.32 and an RMSE of 3.42. Within South Africa, RF has been used for the mapping of soil organic carbon stock predictions by Kotzé and van Tol (2023) using derivatives from environmental covariates such as slope, DEM and a multispectral scanner. It was also used recently by Venter *et al.* (2021) to map soil organic carbon stocks using several decades of satellite imagery over the entire country. On field scale in South Africa, RF was used by Kock *et al.* (2024) with mid-infrared spectroscopy to predict soil chemical properties such as pH, P, K, Ca, Mg and Na in the western highveld region. Kock *et al.* (2024) achieved the most accurate result for Ca (R^2: 0.71, RMSE: 129.06) and Mg (R^2: 0.71, RMSE: 36.75).

2.4.3.3 Fuzzy K-means clustering

To improve the performance of prediction models based on spatial observations, clustering techniques can be used. These techniques work by grouping sets of data that share similar characteristics through various methods. An increasingly popular method for this is Fuzzy K-means clustering (FKM) (Heil et al., 2019). FKM is a soft clustering method as opposed to a hard clustering method, such as K-means clustering (KM). This means that instead of objects belonging to exclusive clusters based on geospatial similarities, they possess of a degree of belonging to several clusters (Hot & Popović-Bugarin, 2015). The degree of belonging of each object adds up to 1, allowing for gradual changes in the soil landscape to be expressed (Hanesch et al., 2001). Because of this, FKM has been found to outperform KM in DSM (Goktepe et al., 2005). FKM has been effectively used by Hot and Popović-Bugarin (2015) to cluster soil chemical data based on soil types in Montenegro and by Heil et al. (2019) to zone soil texture, pH and cation exchange capacity (CEC) in the West African savanna using mid-infrared DRIFT soil spectra. It was also found by Zhao et al. (2023) that FKM could be used to capture the gradual changes in the soil environment and doing so, regionalise large sets of soil property data in China.

2.5 Model validation statistics

2.5.1 Introduction

Several statistical methods used to compare the accuracy of DSM and ML techniques are described by Khaledian and Miller (2020).The most commonly used validation statistics reported by Khaledian and Miller (2020) are the coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE). In this study RMSE, coefficient of variability (Cv), R^2, and Lin's concordance correlation coefficient (CCC) was used. These methods allow for the comparison of predictions to the observed values as found by laboratory analysis in terms of

variation, accuracy and possible bias. Khaledian and Miller (2020) advocates for the use of CCC alongside other accuracy metrics, for its ability to detect bias in the prediction model.

2.5.2 RMSE

Root mean square error is an accuracy metric used to describe how close predicted values are to actual values (Khaledian & Miller, 2020). The RMSE can be determined with Equation 2.5.1 as presented by Dong *et al.* (2018):

Equation 2.5.1: Root mean square error

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

In this equation, n is the sample size, y_j is the observed value and $\hat{y}_j$ is the predicted value.

2.5.3 Cv

The Cv gives RMSE as a percentage of the mean of observations, allowing for evaluation between sets of data Swanepoel *et al.* (2018).

Equation 2.5.2: Coefficient of variability

$$Cv = \frac{RMSE}{Mean} \times 100$$

2.5.4 R^2

R^2 is a metric used to describe variance explained, indicating the proportional variance of a variable represented in Equation 2.5.3 as presented by (Khaledian & Miller, 2020; Dong *et al.,* 2018 and Dong *et al.,* 2019):

Equation 2.5.3: Coefficient of determination

$$R^2 = \frac{\sum_{j=1}^{n}(x_j y_{j-\overline{xy}})^2}{(\sum_{j=1}^{n} x_j^2 - \bar{x}^2)(\sum_{j=1}^{n} y_j^2 - \bar{y}^2)}$$

$x_j \; And \; y_j$ represent the observed and predicted values, respectively, and $\overline{x}, \overline{y}$ are the respective means.

2.5.5 CCC

CCC assess the agreement between observed and predicted values by assessing both accuracy (difference between observed and predicted values as the spread of both around a perfect 1:1 regression line) and precision. According to Khaledian & Miller (2020), this statistical method can handle bias within prediction variance, and give the equation as:

Equation 2.5.4: Lin's concordance correlation coefficient

$$CCC = \frac{2p\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x + \mu_y)^2}$$

In Equation 2.5.4, p is the Pearson correlation coefficient between measurement's variances ($\sigma_x and\ \sigma_y$) and the means of the two measurement methods ($\mu_x and \mu_y$). The values of CCC range from -1 to 1. With 0 being no correlation and 1 being a perfect correlation to the 1:1 line, making -1 a perfect inverse correlation (Khaledian & Miller, 2020).

CHAPTER 3 MATERIALS AND METHODS

3.1 Data acquisition and study site

3.1.1 Study Site

Two fields, Field 1 and Field 2, (Figure 3.1 c and b) were selected on a farm in the North West province of South Africa. The fields were selected because of the availability of soil chemical- and yield- data as well as the variance in soil composition. The site is located between the towns of Sannieshof and Coligny (Figure 3.1a). Field 1 is approximately three and a half kilometres south-east of Field 2.

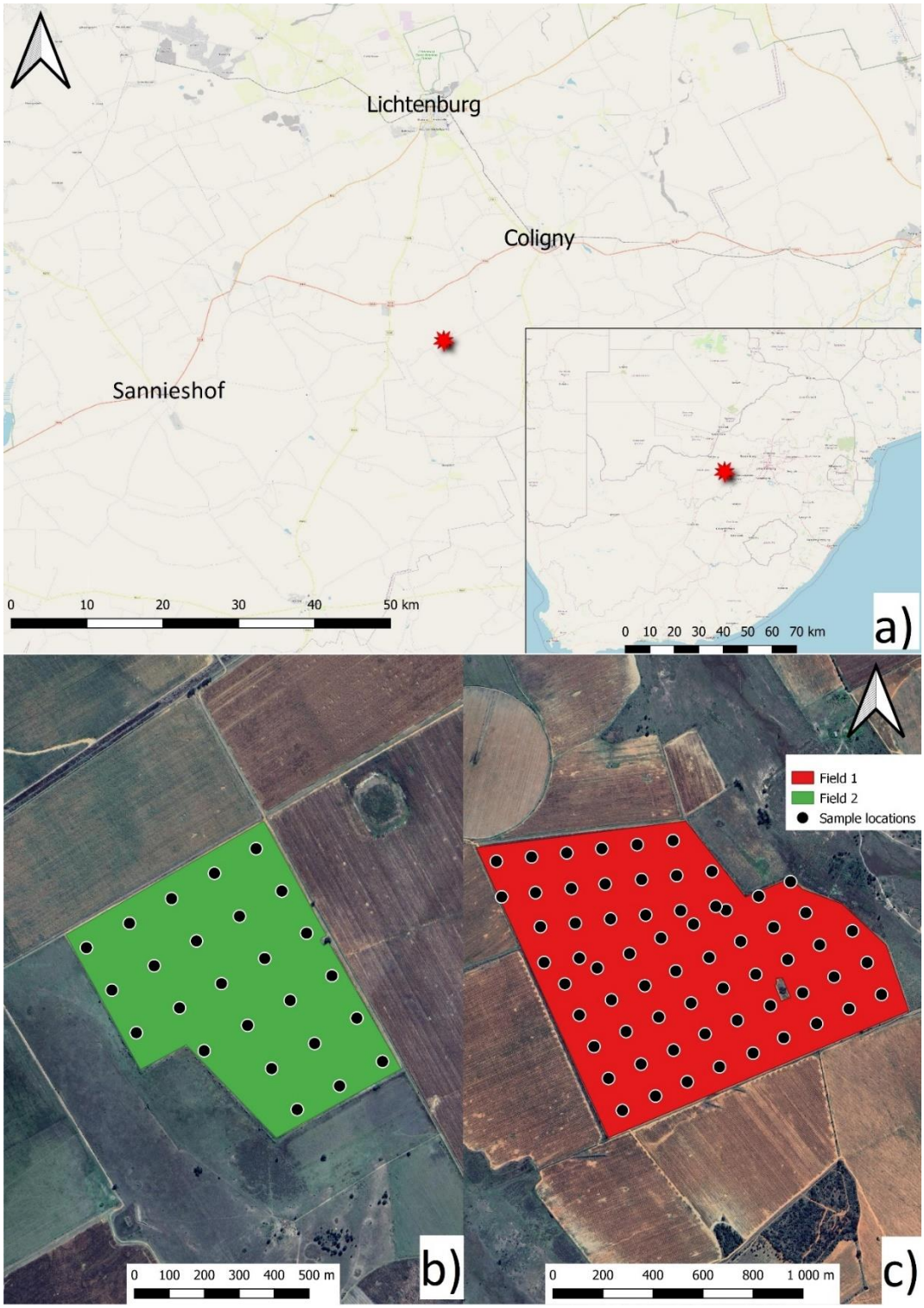

Figure 3.1: The location of the study site (a), showing Field 1 (c) and Field 2 (b).

Field 1 (Figure 3.1c) is approximately 126 ha. The field falls within the Vaal-Vet sandy grassland bioregion of South Africa, characterised by plains with scattered hills (Mucina & Rutherford, 2006). The main geology of Field 1 is undifferentiated tonalite, granite and gneiss (Council of Geoscience, 2007). The broad land type of Field one is Ba, which contains red (<33% of the land

type) and yellow eutrophic, apedal soils with plinthic subsoils (>10% of the land type) (Land Type Survey Staff, 1972-2002). Regarding the observation points in Field 1, several points in the centre of the field are closer than a 2-ha grid, and the orientation of the grid also differs from north to south of the field. This is because, during sampling done by the responsible company, the field was treated as two separate fields, split by a thin road. The treatment and agricultural practices of these two fields were, however, the same for the period regarded by this study.

The smaller Field 2 (Figure 3.1.b) is approximately 44 ha, falling within the Western highveld sandy grassland bioregion, characterised by a flatter terrain with smooth hills and interspersed woody plant species amongst grass (Mucina & Rutherford, 2006). This field covers two geologies, namely the Goedgenoeg and Rietgat formations (Council of Geoscience, 2007). The broad land type of this field is the Bd broad land type, consisting of red (<33% of the land type) and yellow eutrophic, apedal soils with plinthic subsoils (>10% of the land type) (Land Type Survey Staff, 1972-2002).

The historical land use and climate for the two fields are the same. Both fields are planted with maize, getting rotated with a single season of sugar beans every fourth year. Minimum tillage is performed by the farmer, consisting of ripping and cutting before planting. During planting, liquid fertiliser is applied, and after planting, a top dressing is applied. After planting, herbicides are used for weed control. The climate consists of warm summer temperatures (30 ˚C) and very low winter temperatures (sub-0˚C) with severe frost occurring (mean frost days: 37) during wintertime (Schulze et al., 2007). The median annual precipitation is 530 mm, which falls mostly in summer, as the area falls within the summer rainfall area according to Schulze et al. (2007). The mean average temperature is 16.4 ˚C, the mean average potential evaporation is 2423 mm, and the mean average soil moisture stress is 79 % (Mucina & Rutherford, 2006). The driest and coldest time in the region is during June and July, and the hottest, wettest is from December to January.

3.1.2 Data acquisition

3.1.2.1 Soil Data

The data from soil chemical analysis were obtained from NWK for Fields 1 & 2, for 2021. The data represented chemical properties measured from topsoil samples, taken from the top 300 mm of soil with a soil auger, over a 2-ha grid pattern. At each point, five samples are taken across the row of maize; the locations of these samples can be seen in Figure 3.1. The chemical properties included, amongst others, pH, measured in 1:2.5 1M KCl solution, available phosphorus (P)

measured with the Bray II method, and 1M NH4OAc extractable base cations including Calcium (Ca), magnesium (Mg), potassium (K) and sodium (Na). Additionally, soil classification data on a 1-ha grid were obtained for the two fields from NWK. The data included the soil form classified on a one-hectare grid with the 2018 soil classification system (Soil Classification Working Group, 2018) and a soil form map created by NWK. The soil form maps of NWK is drawn by the soil scientist in charge of the classification based on the guidelines of Verster *et al.* (2022). This soil form map was then redrawn in QGIS (QGIS Development Team, 2024) with the same boundaries and transformed into a raster using the rasterize tool. A summary of the main soil forms from this effort appears in Table 3.1.

Table 3.1: Summary of soil forms reported and their horizon sequence

Summary of soil forms reported and their horizon sequence

Soil form	Abbreviation	Topsoil	Subsoil 1	Subsoil 2
Clovelly	Cv	Orthic	Yellow-brown apedal	Lithic
Pinedene	Pn	Orthic	Yellow-brown apedal	Gleyic
Lichtenburg	Lc	Orthic	Red Apedal	Hard plinthic
Hutton	Hu	Orthic	Red Apedal (thick)	

As can be seen in Figure 3.2, Field 1 was dominated by the Hutton soil form with two relatively small pockets of the Lichtenburg soil form in north-west and east of the field. The dominant soil form in Field 2 is the Clovelly form that runs through the centre of the map. The Hutton form then dominates the eastern part of the field while the south, and parts of the west of the field are characterized by pockets of the Pinedene soil form.

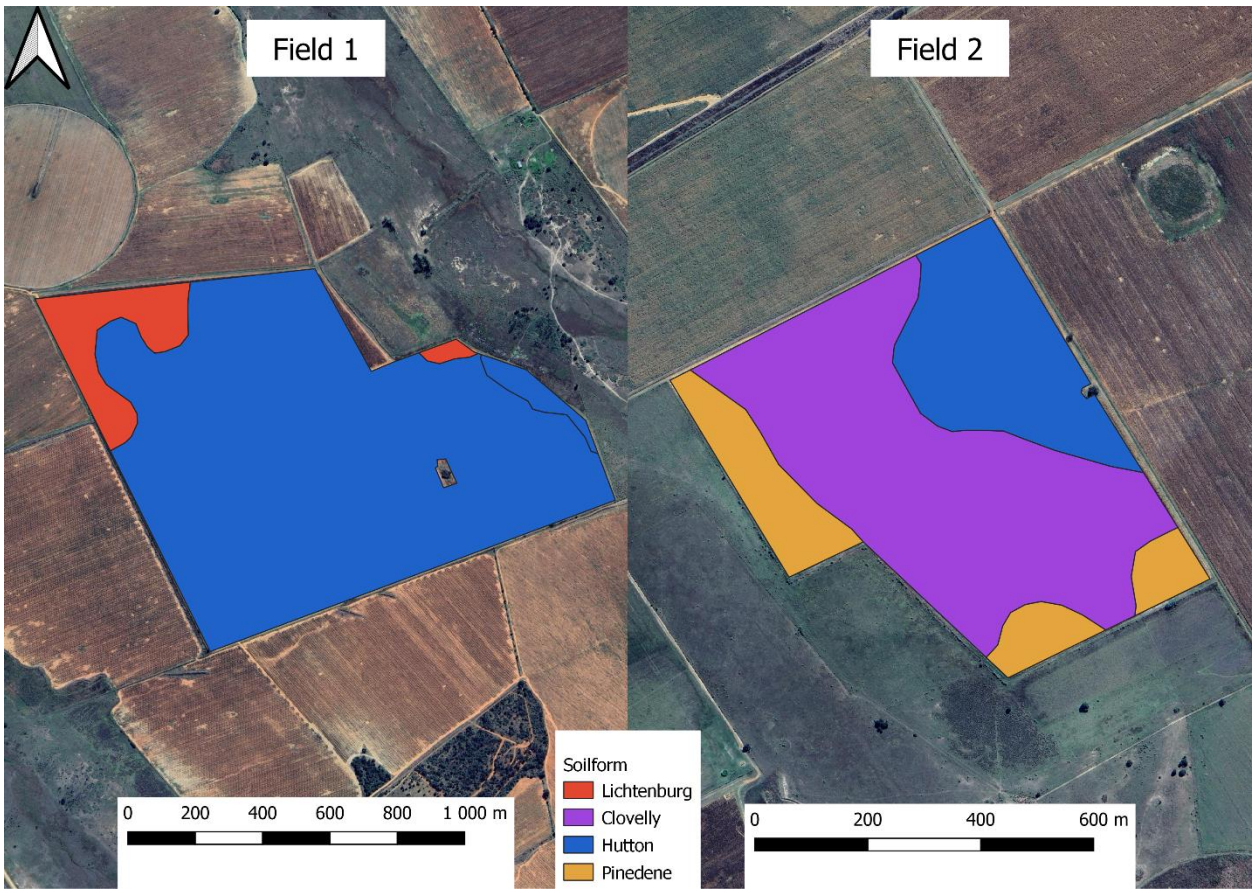

Figure 3.2: Soil forms in Field 1 and 2.

3.1.2.2 Yield data

Yield data for both fields were obtained for 2019, 2020 and 2021. The yield data was obtained from the farmer. Yield was measured through a yield monitor and saved as point data containing, amongst others, dry yield mass and the location where this mass was recorded. Harvesting usually takes two months, starting in July and ending at the end of August. In 2021, large areas of the harvest map were unpopulated as sugar beans were being harvested. As such, the harvest was not included in the covariate set. The Landsat data from this period was still used, as it coincides with the chemical grid sampling done in that year, before sugar beans were planted.

3.1.2.3 Covariate data

Spectral data was collected from Sentinel-2 satellite imagery from the EO browser (Sentinel Hub, 2024) for 2018, 2019, 2020 and 2021. The imagery was chosen to include both a preharvest (Wet) (January and February) and postharvest (Dry) (November) season with minimal cloud coverage; the exact dates are shown in Table 3.2.

Table 3.2: Landsat 2 spectral imagery dates

Field 1 and 2 Landsat 2 imagery dates

Season	Date
Wet	2018/01/25
Dry	2018/11/12
Wet	2019/01/25
Dry	2019/11/06
Wet	2020/01/25
Dry	2020/11/15
Wet	2021/02/13
Dry	2021/11/15

The bands 2, 3, 4 and 8, representing the blue (B), green (G), red (R) and infrared (IR) bands, respectively, were downloaded. The following spectral indices were calculated: Redness index (RI), brightness index (BI), colouration index (CI) (Flynn et al., 2022), and normalised difference vegetation index (NDVI; Rousse et al., 1973) using Equations 3.1.1 to 3.1.4. This was done using the raster calculator tool SAGA GIS 2.3.2 (Conrad et al., 2015) and exported to TIFF files.

Equation 3.1.1: Normalised difference vegetation index

$$NDVI = \frac{NIR-R}{NIR+R}$$

Equation 3.1.2: Redness index

$$RI = \frac{R^2}{B \times G^3}$$

Equation 3.1.3: Brightness index

$$BI = \frac{(R^2+G^2+B^2)}{3^{0.5}}$$

Equation 3.1.4: Colouration index

$$CI = \frac{R-G}{R+G}$$

For elevation, the 30 m the shuttle radar topography mission (SRTM) digital elevation model (DEM) was obtained from Earth Explorer (USGS, 2024). Within SAGA 2.3.2 (Conrad et al., 2015) topographic derivatives were calculated with the Basic terrain analysis tool. A summary of the covariate types and the detailed derivatives used can be found in Table 3.3.

Table 3.3: Summary of covariates used for the creation of Random Forest and Cubist prediction models

Covariates used in the creation of Cubist and Random Forest prediction models		
Covariate Type	**Details**	**Years/States**
Landsat-derived indices	Colouration index (CI), Brightness index (BI), Redness index (RI) and Normalized Difference Vegetation Index (NDVI)	2018, 2019, 2020 & 2021. Wet and dry
Digital Elevation Model	Derived: Elevation, Analytical hill shading, Slope, Aspect, Cross-Sectional Curvature, Longitudinal curvature, Convergence index, Flow Accumulation, Topographic Wetness Index, LS Factor, Channel Network Base level, Vertical Distance to Channel Network, Valley Depth and Relative slope position	Not applicable
Yield data	Yield data for the field	2019 and 2020
Soil form classification	Soil classification based on soil classification guidelines (Soil Classification Working Group, 2018) is used as a factor.	Not applicable

All data layers were reprojected to WGS 84 UTM 35S, to enable distance analyses. The soil maps were rasterised in QGIS (QGIS Development Team, 2024), and all covariate layers were resampled to fit onto the same 20 m by 20 m pixel resolution in R (R Core Team, 2023) during the raster stacking process.

A selection of covariates used for Field 1 appears in Figure 3.3 and in Figure 3.4 for Field 2. It should be noted that these are the raw rasters before they were resampled during the previously mentioned step. As can be seen in the yield maps for both Fields in 2019 and 2020, a higher yield was recorded for 2020. These yield maps also coincide with the NDVI maps. Field 1 had a greater difference in slope than Field 2. The higher yield in 2020 is likely because of the higher rainfall recorded at the time. Oppaslaagte Grain Silo, the closest silo from both fields (4 kilometres)

reported 394 mm for the period between July 2018 and June 2019; 756 mm between June 2019 and June 2020; and 494 mm between July 2020 and June 2021.

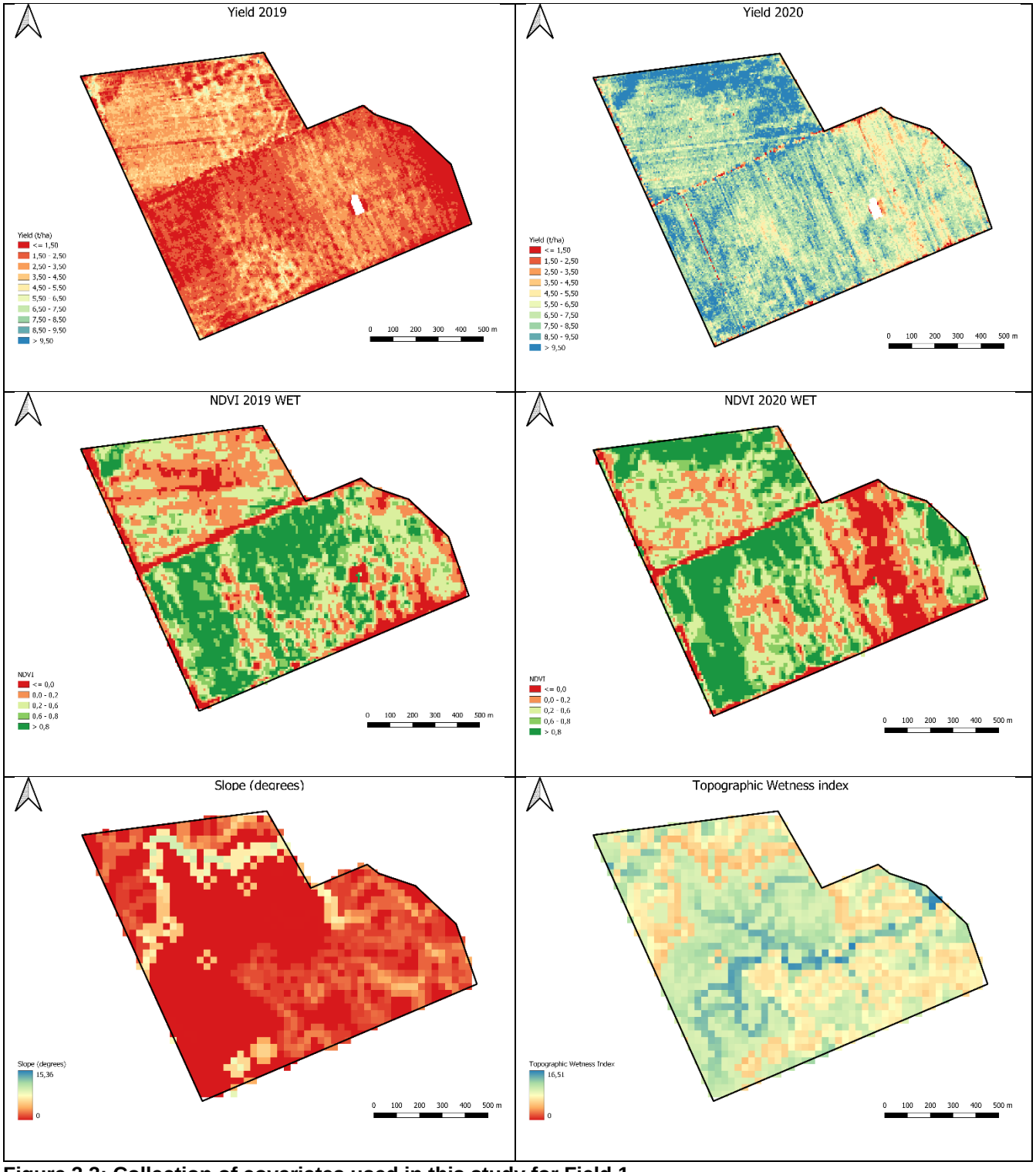

Figure 3.3: Collection of covariates used in this study for Field 1.

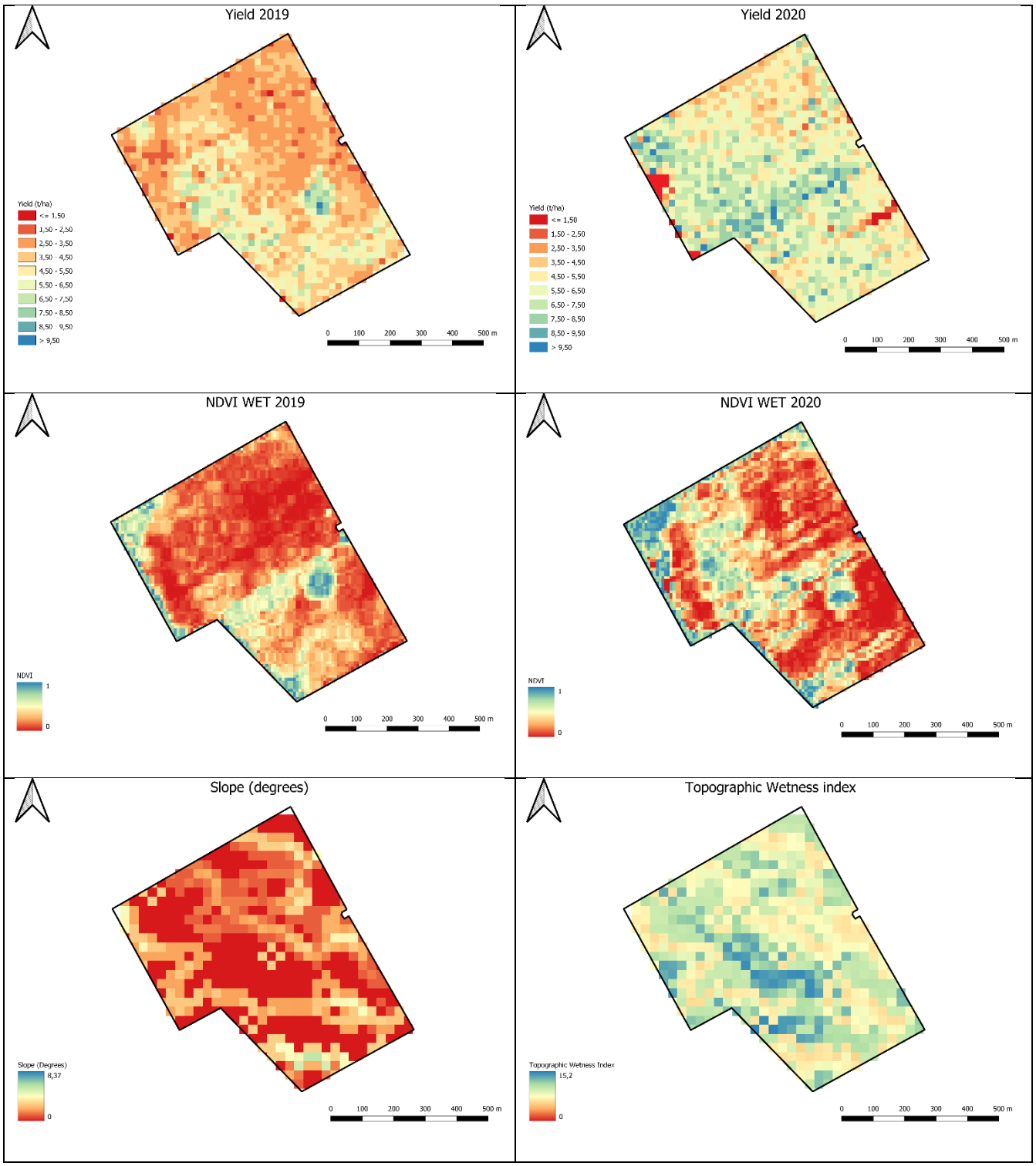

Figure 3.4: Collection of covariates used in this study for Field 2.

3.2 Conventional maps (OK)

A set of soil property maps for each field was created using the chemical analysis done on a 2-ha grid, as set out in Figure 3.1. To do this, the observation points and study site area were imported to R (R Core team, 2023). A local empirical- and theoretical- semi-variogram model was

then created for K, P, pH, Ca, Mg and Na, respectively, using the gstat package (Pebesma, E.J., 2004 and Gräler et al., 2016) and fitted. Using the stats package (R Core Team, 2023), OK was then used to create an interpolation map for each soil property based on the variogram models. The resulting maps were mapped with a 20 m resolution. The legend of each map was classified according to an interpretation of the Fertiliser Recommendation manual of South Africa's (FERTASA, 2007) fertiliser recommendation guidelines. These map classes are shown in Table 3.4.

Table 3.4: Map classes for maps representing soil chemical properties based on FERTASA (2007) guidelines

Table of map classes for maps representing soil attributes based on FERTASA (2007)					
Map	**Class**	**Interpretation of class**	**Map**	**Class**	**Interpretation of class**
K (mg/kg)	0-40	Below optimal	**Ca (mg/kg)**	0-50	Below optimal
	40-60			50-100	
	60-80			100-150	
	80-120			150-250	
	120-180	Optimal		250-350	Optimal
	180-220			350-500	
	220-300			500-700	
	300-400	Above optimal		700-1000	Above optimal
	400-600			1000-1800	
	600-800			1800-3000	
				3000-5000	
P (mg/kg)	0-4	Below optimal	**Mg (mg/kg)**	0-20	Below optimal
	4-8			20-40	
	8-14			40-60	
	14-21			60-80	
	21-30	Optimal		80-150	Optimal
	30-35	Above optimal		150-300	Above optimal
	35-60			300-400	
	60-90			400-500	
	90-200			500-700	
				700-800	
				800-2000	
pH	3,5-4	Below optimal	**Na (mg/kg)**	0-10	Optimal range
	4,0-4,5			10-15	
	4,5-5,0	Optimal		15-20	Above optimal
	5,0-5,5			20-30	
	5,5-6,0			30-40	
	6,0-6,5			40-60	
	6,5-7,0	Above optimal		60-90	
	7,0-7,5			90-120	
	7,5-8,0			120-220	

3.3 Machine learning DSM

3.3.1 Creation of covariate data frame

To implement ML DSM techniques, a database needs to be created of covariate data. To do this, the soil chemical data was imported to R as a shp file. Next, all the spectral, yield and topographical data (in raster format) were imported into R. The soil form maps were then imported, and the soil forms were defined as a factor. The collection of rasters was then resampled, using the terra package (Hijmans, 2023), to the same 20 m grid and stacked with the utils package (R Core Team, 2023). This raster stack was then extracted to the same locations as the soil samples and combined with the soil properties data to form a combined DSM data stack. This was done with the terra package (Hijmans, 2023). Lastly, this DSM data stack was cleaned of missing values with the stats package (R Core Team, 2023).

3.3.2 Creation of ML Models

The DSM data stack was split between training (70%) and validation (30%) sets using Fuzzy K-means clustering, done with the ppclust package (Cebeci, 2018). For each soil property at each field, both Cubist and Random Forest ML algorithms were created and applied using the Cubist (Kuhn and Quinlan, 2023) and randomForest (Liaw and Wiener, 2002) packages, respectively, with the training dataset. The caret package (Kuhn, 2008) was then used to fine-tune the parameters of the Cubist and Random Forest models. The soil property maps were made by applying the Cubist and Random Forests algorithms to the covariate stacks of the two sites.

3.4 Model Validation

3.4.1 Ordinary Kriging

The OK model was assessed using a leave-one-out cross-validation, with the gstat package. The R^2, RMSE and CCC was calculated as validation metrics. Further validation was then done through the interpretation of the variogram statistics and a visual inspection of the variogram for each field and soil property. Lastly, the observation points were plotted on top of the OK maps, using the same class intervals to visualise differences between observations and predicted maps. This was done for each field and soil property.

3.4.2 Cubist and Random Forest

To validate the Cubist and Random Forest models, the goodness of fit was measured using the ithir package (Malone, 2021), making use of the validation data set. The statistics calculated for validation include R^2, RMSE and CCC. The ML method with the best-performing validation was then compared to OK. As with the OK maps, the observations were mapped with the same class intervals on top of the ML prediction map for visual inspection of the correlation between observed values and predictions.

3.4.3 Comparison limitations

Two different validation techniques are used for OK and the ML DSM techniques because of the nature of these predictive techniques. Optimally, an independent data set consisting of observations in between the existing points would be used, but as legacy data is used for this study, such a data set could not be collected. Also, because of the large grid resolution, removing any observations would drastically influence die forming of the semivariogram for OK, as it is very sensitive to inter-sampling range (ZHU & Lin, 2010). The ML DSM techniques are, however, more sensitive to how well the observations coincide with specific environmental factors (Khaledian & Miller, 2020), allowing for data partitioning where the training and validation sets are kept separate when the model is created. As such, the comparison done between OK and ML DSM isn't a direct comparison.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Descriptive statistics

4.1.1 Yield results

A summary of the yield maps for 2019 and 2020 is shown in Figure 4.1. The highest mean yields for both fields were obtained in 2020 (Field 1: 7.58 t/ha and Field 2: 6.15 t/ha), and the lowest mean yields were obtained in 2019 (Field 1: 2.53 t/ha and Field 2: 4.19 t/ha). This is below the annual yields for white maize (5.33 to 7.25 t/ha) in South Africa under dry land production (Department of Agriculture, Land Reform and Rural Development, 2022), at its lowest in 2019 but within the average for 2020. Both fields showed a similar range in 2019, with a range of 16.08 t/ha in Field 1 and 16.64 t/ha in Field 2. In 2020, Field 1 showed a higher yield range with a range of 23.97 t/ha than Field 2 with a range of 20.33 t/ha. The higher yield is most likely the result of the higher rainfall in that year. The NWK Oppaslaagte grain silo, within 10 kilometres of the study

sites, reported 394 mm of rain for the period between July 2018 and June 2019, 756 mm between July 2019 and June 2020 and finally 494 mm between July 2020 and June 2021.

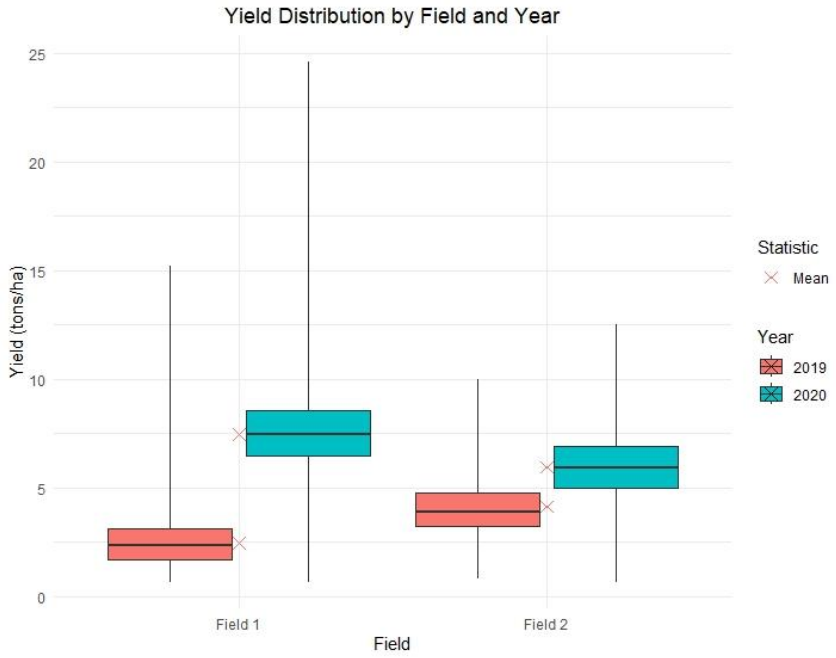

Figure 4.1: Yield results (tons/ha) for Fields 1 and 2 for the 2019 and 2020 harvest seasons.

4.1.2 Summary statistics of observations and results

Figures 4.2 and 4.3 show the summary statistics of the observations and results for Fields 1 and 2. The observed values are the results of the chemical analysis done on each point as set out in Figure 3.1, while the values of OK, Cubist and RF derive from the predicted values of each pixel in the prediction maps created by that method. The figures help present the ability of each method to predict the distribution of different soil properties and give an overview of how well they approximated the observed values of each respective soil property. A common occurrence in all these plots is that the prediction method plots have higher minimum and lower maximum values than that of the observed values, with the only exceptions being the Cubist prediction of Na in Field 1 (Figure 4.2 f) and Field 2 the OK prediction of pH and the Cubist prediction of K (Figure 4.3 a and c). This shows a general smoothing of the data, especially at the lower values, as the minimum values for the predicted maps are higher than the measured data. Also of note is that in the case of P for Field 1 (Figure 4.2a), the box and whisker plots of the Cubist and RF methods consist of a single line; this is also present in the plot for the Cubist method for Mg in Field 2 (Figure 4.3e). This is indicative of that method failing at capturing spatial variation and just creating a map filled with an average value. This will be further explored in the following sections.

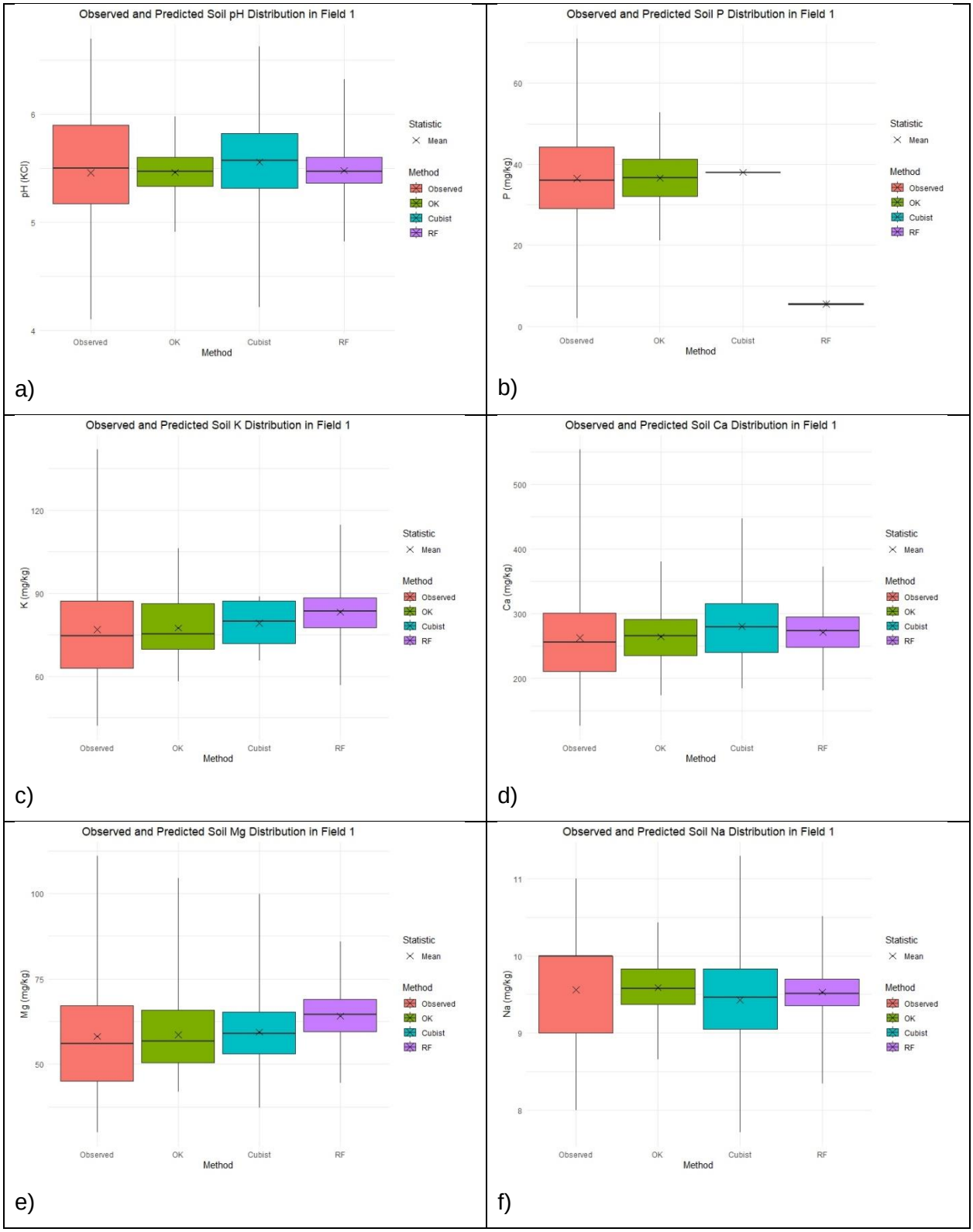

Figure 4.2: Box and Whisker Plots of Summary Statistics for Field 1 for soil pH (a), P (b), K (c), Ca (d), Mg (e) and Na (f) for each method (Observed, Ordinary Kriging (OK), Cubist and Random Forest (RF).

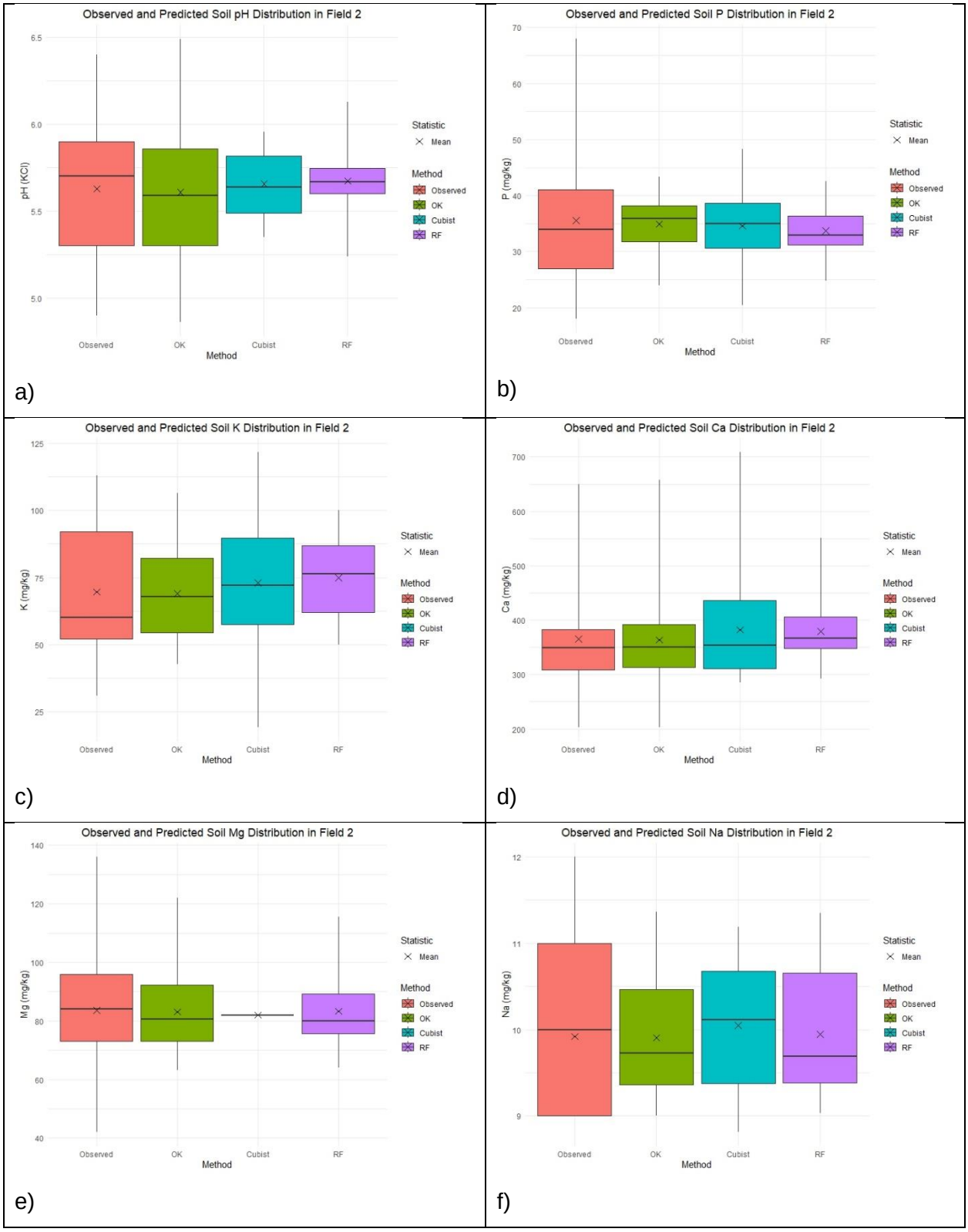

Figure 4.3: Box and Whisker Plots of Summary Statistics for Field 2 for soil pH (a), P (b), K (c), Ca (d), Mg (e) and Na (f) for each method (Observed, Ordinary Kriging (OK), Cubist and Random Forest (RF).

The limitations of OK and ML DSM techniques are therefore expressed in the general increase in minimum values and reduction in maximum values, resulting in smaller ranges throughout. This

smoothing effect was less for OK than for ML DSM. A likely explanation is that OK was done with all data points, while RF and Cubist models only used 70 % of the points during training. Thus, minimum and maximum values can be excluded from training data. The smoothing effect of both OK, described by Yamamoto (2005), and ML (Khaledian & Miller, 2020), can be seen in the smaller ranges and minimum values. The results and performance of each prediction method are explored in full in the following chapters.

4.2 Validation of Conventional maps created using Ordinary Kriging and Machine learning digital soil mapping techniques

In general, neither the OK nor the ML DSM methods were able to create sufficiently accurate maps of soil properties for both Field 1 and Field 2 (Tables 4.1 and 4.2). In Field 1, only moderately accurate validation results were obtained through OK for K and Mg, while poor accuracy was achieved for the rest of the properties (Table 4.1). In Table 4.2, the validation results for OK in Field 2 show better results, with only P and Na showing poor correlation and accuracy. The performance of each method for the different soil properties is discussed in full below.

Table 4.1: Validation statistics of Cubist, Random Forest and Ordinary Kriging prediction techniques for Field 1

Validation measure	Property								
	pH			**P**			**K**		
Model	**Cub**	**RF**	**OK**	**Cub**	**RF**	**OK**	**Cub**	**RF**	**OK**
R^2	0,11	0,11	-0,12	NA	0,01	0,00	0,01	0,01	0,38
CCC	-0,27	-0,16	-0,03	0	0,02	0,00	0,04	-0,03	0,25
RMSE	0,81	0,68	0,51	14,95	14,82	13,28	18,93	23,46	18,07
	Ca			**Mg**			**Na**		
Model	**Cub**	**RF**	**OK**	**Cub**	**RF**	**OK**	**Cub**	**RF**	**OK**
R^2	0,06	0,16	0,14	0,03	0,00	0,49	0,02	0,03	-0,03
CCC	-0,18	-0,25	0,08	-0,09	-0,01	0,38	-0,11	-0,04	-0,01
RMSE	115,76	111,95	74,21	18,13	21,16	14,73	1,22	0,92	0,79

OK = Ordinary Kriging; Rf = Random Forest; Cub = Cubist

Table 4.2: Validation statistics of Cubist, Random Forest and Ordinary Kriging prediction techniques for Field 2

Validation measure	Property								
	pH			**P**			**K**		
Model	**Cub**	**RF**	**OK**	**Cub**	**RF**	**OK**	**Cub**	**RF**	**OK**
R^2	0,35	0,36	0,58	0,17	0,16	-0,08	0,69	0,09	0,60
CCC	0,42	0,25	0,53	0,23	0,07	-0,03	-0,56	-0,18	0,56
RMSE	0,29	0,26	0,32	12,85	14,13	11,34	37,17	25,53	19,24
	Ca			**Mg**			**Na**		
Model	**Cub**	**RF**	**OK**	**Cub**	**RF**	**OK**	**Cub**	**RF**	**OK**
R^2	0,44	0,10	0,66	NA	0,26	0,63	0,50	0,81	0,48
CCC	0,50	0,12	0,52	0,00	0,20	0,57	0,46	0,66	0,33
RMSE	55,94	78,70	75,21	21,04	19,16	15,64	0,84	0,60	0,86

OK = Ordinary Kriging; Rf = Random Forest; Cub = Cubist

4.2.1 pH

Poor validation results were achieved for OK pH predictions in Field 1 (Table 4.1), with both R^2 and CCC being below zero (R^2: -0.12 and CCC: -0.03). The low R^2 implies that very little variability in pH was described by the OK model, while the low CCC implies that there exists poor agreement between the observed and predicted values (Khaledian & Miller, 2020). In contrast, the OK model of pH in Field 2 achieved moderately accurate results with an R^2 of 0.58 and a CCC of 0.53. The R^2 results in Field 2 were higher than those found by ZHU and Lin (2010) (R^2: 0.27) and those of Tziachris *et al.* (2020) (R^2: 0.236). The OK results for pH in Field 1 are thus not accurate enough for use in PA, whilst the results in Field 2 indicate that OK can be used to predict pH based on its

R^2 and CCC values. In terms of RMSE, 0.32 for Field 2 (Table 4.2) is lower than the property class as used in Table 3.4 at the measured mean value of 5.0 - 5.5. This means that the prediction error of OK is smaller than the mapping classes used for fertiliser recommendations in the case of Field 2. The RMSE in Field 1 is higher, but as the R^2 and CCC are very poor, the map of OK is already not usable in PA.

Both Cubist and RF achieved poor performance results in Field 1 (Table 4.1), with Cubist having a poor R^2 of 0.11 and a very poor CCC of -0.27. RF performed slightly better, but still poorly, with an R^2 of 0.11 and CCC of -0.16. This implies that both Cubist and RF explained a similar amount of variance according to their R^2 values, whilst both had poor agreement between observed and predicted values (Khaledian & Miller, 2020). Both the RMSE of Cubist (0.81) and RF (0.68) are higher than the property class used for mapping (Table 3.4) at the measured mean value of 5.63. This implies that the wrong fertiliser recommendation can be made, as the prediction error is in the same order of magnitude as that of the mapping classes. Based on these results, neither the Cubist nor the RF prediction maps for pH in Field 1 could be considered accurate. The ML DSM techniques performed significantly better for pH in Field 2, with Cubist having an R^2 of 0.35 and CCC of 0.42, whilst RF achieved an R^2 of 0.36 and CCC of 0.25 (Table 4.2), making Cubist the best performing ML DSM map for pH in Field 2. The RMSE of both Cubist (0.29) and RF (0.26) is very low when compared to the mean of observed pH in Field 2 and the mapping class, leaving little danger of wrongful remedial action based on prediction error. The R^2 results in Field 2 are similar to those achieved by Suleymanov *et al.* (2023) who found Cubist to have an R^2 of 0.34 and RF to have an R^2 of 0.44 when predicting pH (KCl). Even more similar results for RF predicting soil pH(KCl) were found by Dharumarajan *et al.* (2017) with an R^2 of 0.30 and a CCC of 0.37. The accuracy is however, still low, making the map unusable for PA. In Field 1 neither OK or the best performing ML DSM technique, RF, managed to sufficiently describe and capture the spatial distribution of pH whilst OK best described pH in Field 2 as compared to RF.

The maps created through OK and the ML DSM techniques for pH in Field 1 (Figure 4.4) and Field 2 (Figure 4.5) allow the correlation between the observed values (points) and the prediction maps to be visually assessed. As can be seen in Figure 4.4, like the graphs in Figure 4.2a, the highest and lowest pH values are not represented within the maps, though the Cubist map did represent a wider range of values. There is also a large disconnect between the observed values and each prediction map. The observations between 5.0 and 6.0, which are around the mean value for Field 1, are more closely represented by the prediction maps. The OK map of Field 1 is like the RF map with similar distributions of pH, while the Cubist map showed much more spatial variation.

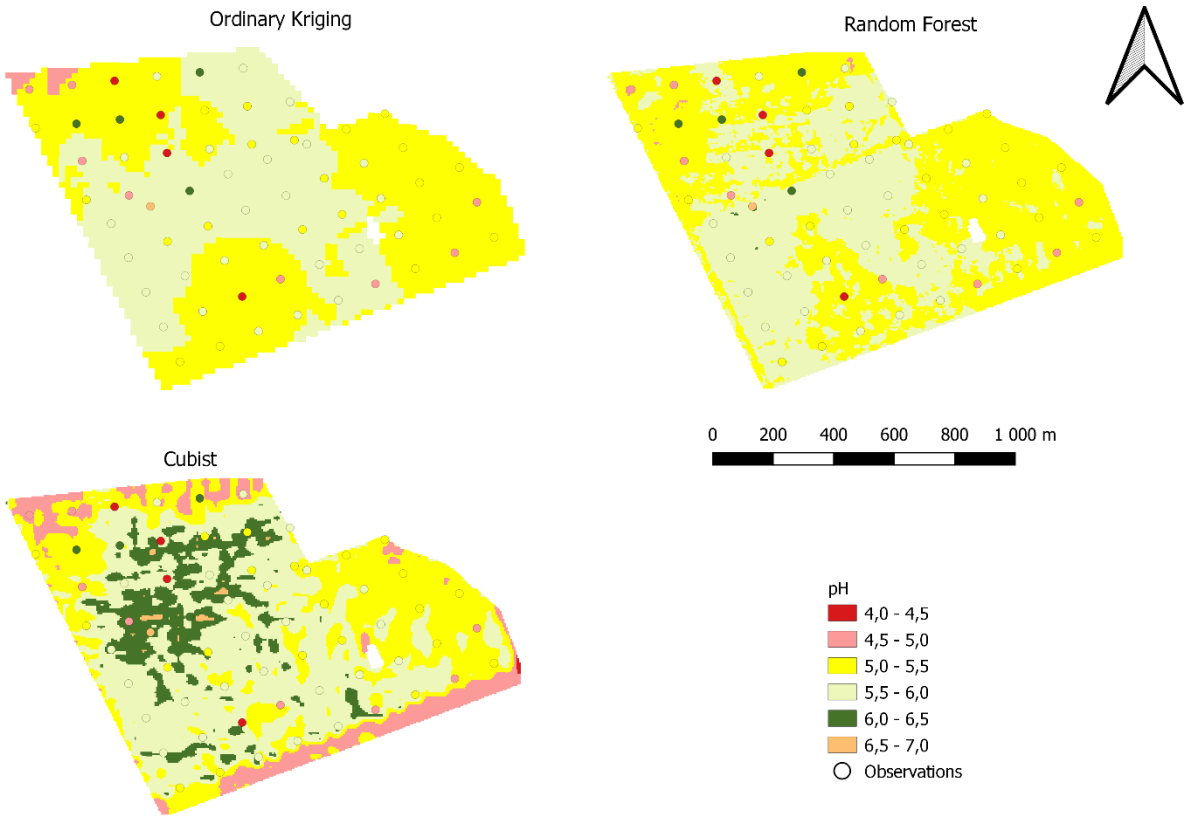

Figure 4.4: OK and ML DSM maps created for pH for Field 1.

In the case of Field 2 (Figure 4.5), the OK map follows the spatial pattern of the observations very closely compared to the ML DSM techniques. The spatial distribution patterns of pH for Cubist and RF are also very similar to each other, although they miss the highest and lowest values of the observations. In this case, a possible explanation for this is that as Field 2 is smaller, with fewer observation points, these high and low values were under-represented in the training data set.

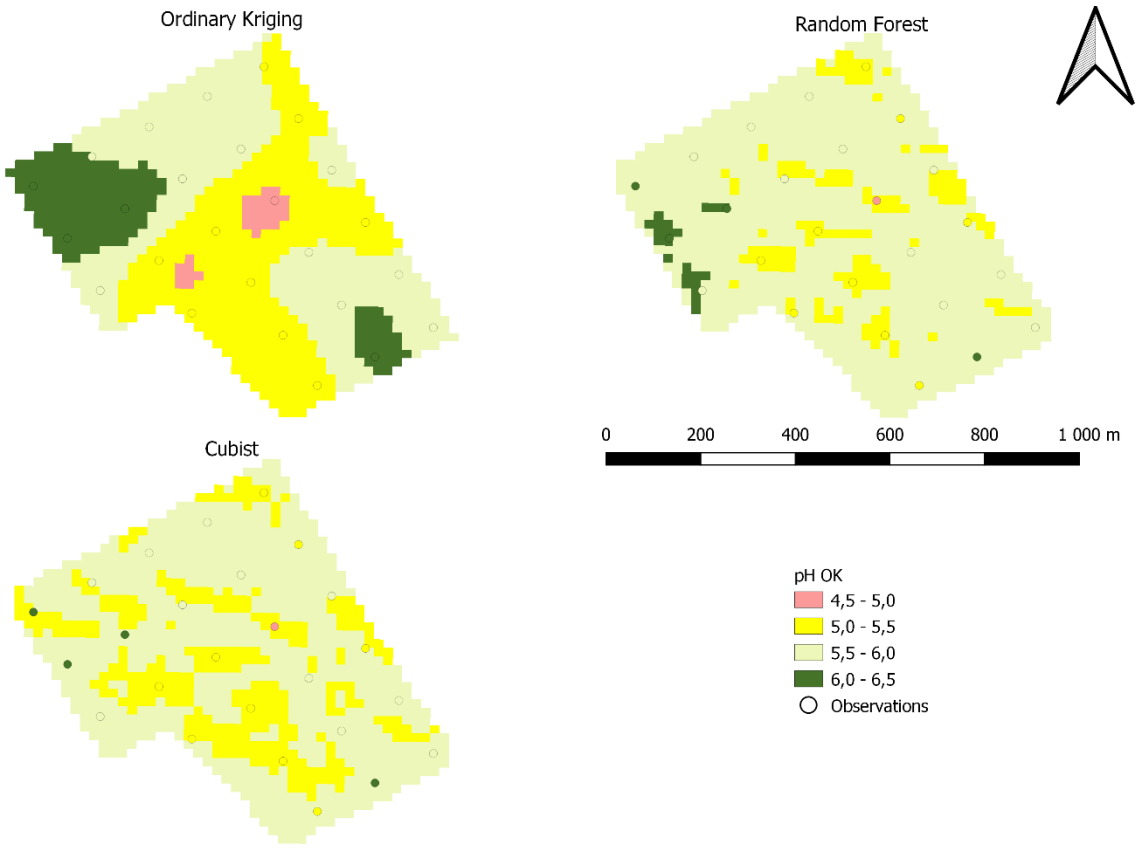

Figure 4.5: OK and ML DSM maps created of pH for Field 2.

4.2.2 P

OK P predictions achieved poor validation results in both Fields 1 and 2, in terms of R^2 (0 and -0.8) and CCC (0 and -0.03) as can be seen in Tables 4.1 and 4.2. In Figures 4.2b and 4.3b, it can also be seen that the range of predicted values is much lower than that of the observations. OK, thus had an extreme smoothing effect on the data in both Fields. The very low R^2 and CCC results in both cases indicate poor capture of variation by OK and little correlation between observed and predicted values (Khaledian & Miller, 2020). Based on these validation results, the OK maps of both Fields are not accurate enough for use in PA. The RMSEs of Both Fields are also relatively high (13.28 mg/kg and 11.34 mg/kg) when compared to the mapping class of 30-35 mg/kg around the mean of the observations in both Fields 1,35.56 mg/kg, and 2, 36.47 mg/kg. These values are, however, close to the optimal P range, 21 to 30 mg/kg, for maize production according to FERTASA (2017) guidelines.

In Field 1, the R^2 value could not be determined for Cubist as the model failed to find rules for the prediction and instead chose a singular value close to the mean of observations. The CCC of this is very poor, -0.27, whilst RF has achieved a poor R^2 of 0.01 and a poor CCC of 0.02, Table 4.1. In the case of RF, the model also failed and instead chose a value close to the minimum with

slight variation throughout the prediction. This means that neither the Cubist nor the RF model could describe the variation of P in Field 1. As these models split data based on feature thresholds and not spatial relationships, it is likely that the training and validation data sets failed to be representative of the existing spatial pattern related to P (Geerts, 2024). Better validation results were achieved by both Cubist and RF in Field 2, Table 4.2, with Cubist achieving an R^2 of 0.17 and CCC of 0.23, whilst RF achieved an R^2 of 0.16 and CCC of 0.07. These R^2 and CCC values were, however, still poor, with Cubist slightly outperforming the RF prediction. The RMSE value of Cubist was 12.85 mg/kg (Table 4.2) and around the mean of the observed of P in Field 2, 36.47 mg/kg, the prediction error would not result in predictions falling in the below optimal mapping class (21-30 mg/kg), thus not having a major impact on fertilizer recommendations. In the case of Cubist in Field 2, better results were achieved than those of Kaya et al. (2022), who achieved a CCC of 0.13 and an RMSE of 15.94 mg/kg. Though the Kaya study was done on a bigger scale of a one- by one- kilometre resolution. In Field 1, neither OK nor RF managed to accurately predict the distribution of P within the Field. In Field 2, OK failed to predict the distribution of P, while Cubist only managed to poorly predict the distribution thereof.

In Field 1 neither OK nor the ML DSM techniques managed to accurately predict the distribution of P within the Field. Although better validation results were achieved In Field 2, OK failed to predict the distribution of P while Cubist only managed to poorly predict the distribution thereof.

The prediction maps of OK and the ML DSM techniques for Field 1 are given in Figure 4.6, whilst the prediction maps of Field 2 are given in Figure 4.7. As the RF and Cubist models failed in Field 1, these maps consist of a single value and have no similarities to the OK map in Figure 4.6. Although there are several low values of the observations not represented in the OK prediction map, the pattern thereof does somewhat follow the distribution of the observations in Field 2. It is, however, still clear that the OK prediction method failed to accurately capture the spatial distribution of P variation in Field 1. Somewhat better results are shown in Figure 4.6, where the spatial distribution of P variation was somewhat captured by all three prediction methods. A similar pattern can also be seen in the maps, with all three predicting the highest values in the eastern area of the map and lower values in the west of Field 2. The only method to capture values in the lowest range (14-21 mg/kg) of the observations was Cubist.

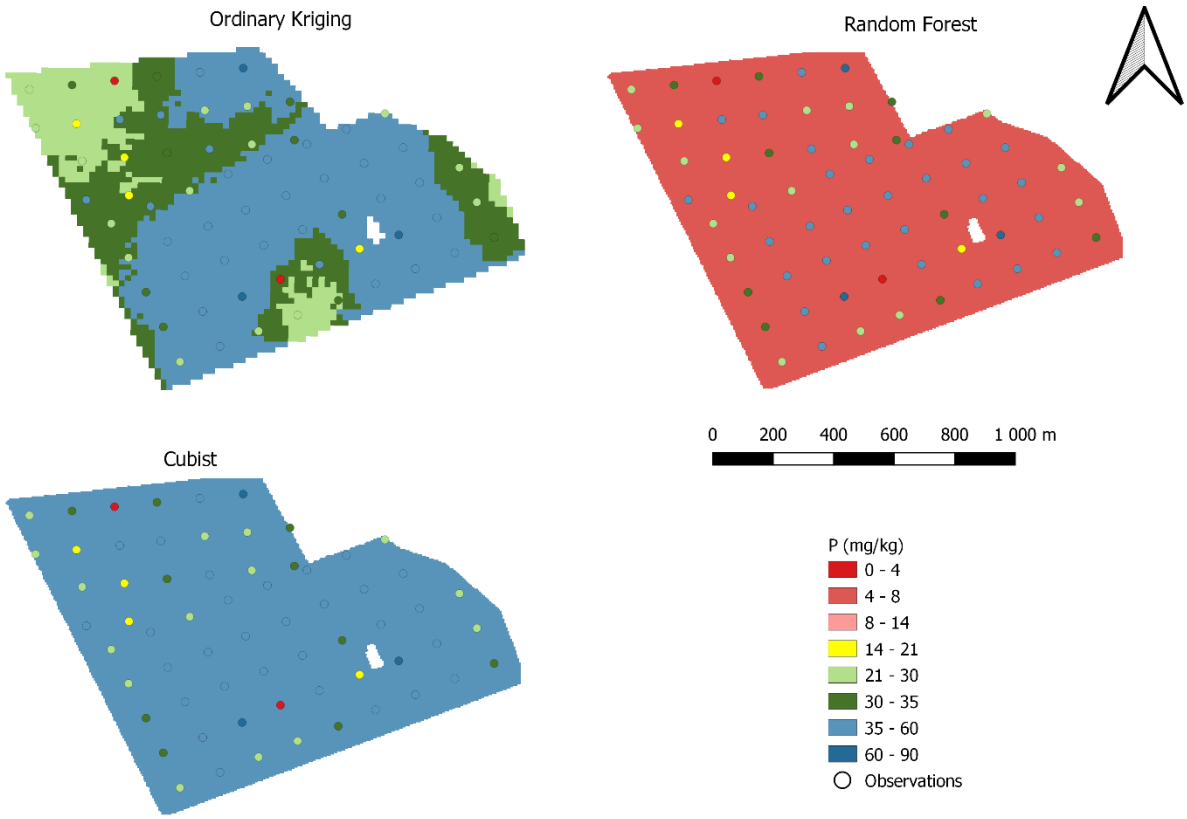

Figure 4.6: OK and ML DSM maps created of P for Field 1

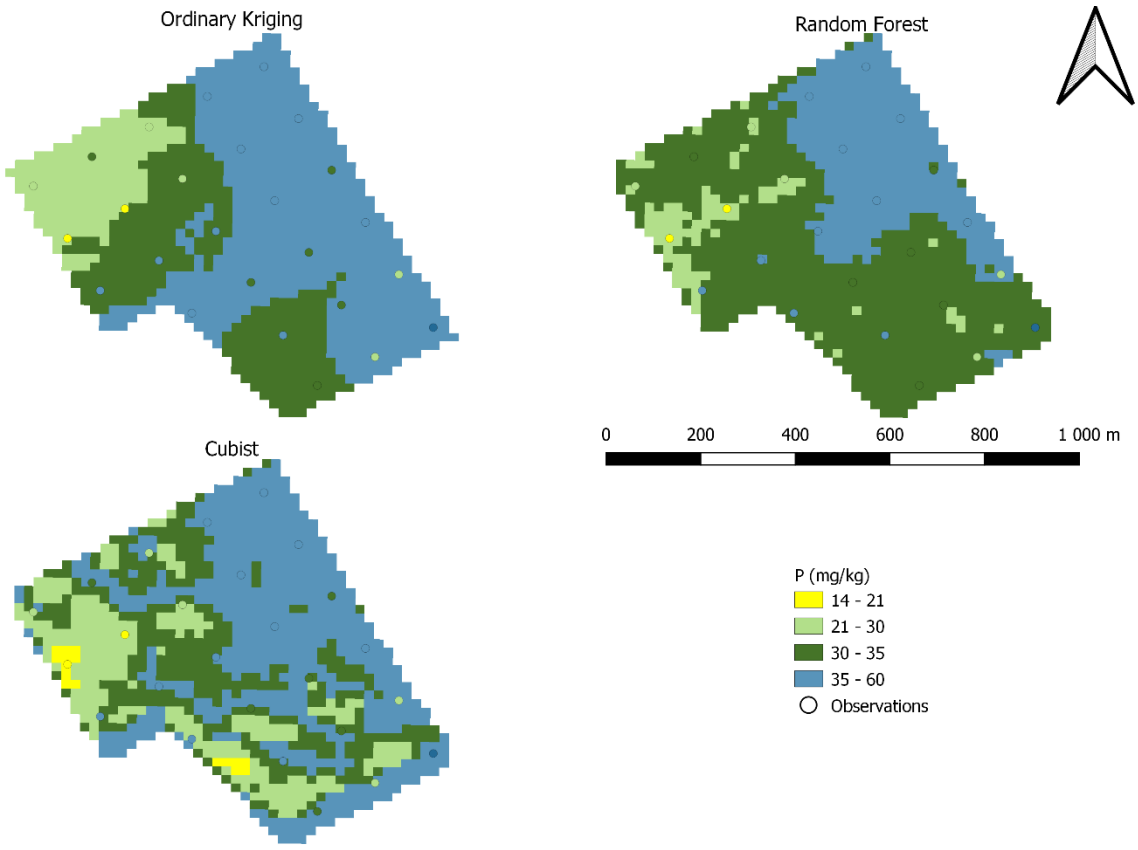

Figure 4.7: OK and ML DSM maps created of P for Field 2.

4.2.3 K

The validation results for OK of K in Field 1 showed poor accuracy with an R^2 of 0.38 and CCC of 0.25 (Table 4.1). This implies a low explanation of variation by the model and poor correlation between predictions and observations (Khaledian & Miller, 2020). The relatively high RMSE of 18.07 mg/kg is larger than the property class as used by FERTASA (2007) at the measured mean value of 60-80 mg/kg. This indicates that the predicted map using OK produces an error in the same order of magnitude as the mapping classes, whereby fertiliser recommendations are made. Therefore, it is quite conceivable that wrong fertiliser recommendations would result from using the OK map. In Field 2, a good R^2 of 0.60 was achieved alongside a moderate CCC of 0.56 (Table 4.2), meaning that the OK model could relatively accurately describe the variation of K and that its predictions could follow the observations. The RMSE of OK in Field 2 is 19.24 mg/kg, which is around the mean of the observations at 96.91 mg/kg, and would still have a small impact on the class used by FERTASA (2007), though less than in the case of Field 1. In terms of the R^2 value

and CCC in Field 2, OK would still be accurate enough for use in PA, though attention should be given to values lower and close to the mean for possible prediction errors.

As can be seen in Table 4.1, the Cubist ML DSM method achieved better validation results compared to RF in Field 1. These validation results were still poor, with an R^2 of 0.01 and CCC of 0.04 in the case of Cubist and a R^2 of 0.01 and CCC of -0.03 for RF. Thus, both models poorly describe the variation of K in field 1 and do not sufficiently agree with the observed values. In Field 2, neither Cubist nor RF performed well. The validations of Cubist resulted in an R^2 of 0.69, which was good, but a CCC of -0.56, which is very poor. The high R^2 value implies that the model captured the variation of K well, while the low CCC implies that there is a moderately negative correlation between the observed and predicted values. The RF model resulted in an R^2 of 0.09, which is poor and a CCC of -0.18 (Table 4.2), meaning this model poorly explained the variation of K and showed poor correlation between observed and predicted values.

In Field 1, neither OK nor the best performing ML DSM technique managed to achieve satisfactory prediction results for K. Even so, OK outperformed the Cubist prediction with a higher R^2 and CCC. In Field 2 OK outperformed the RF model. Both the OK models are accurate enough for use in PA based on the validation results.

The OK and ML DSM prediction maps for Fields 1 & 2 are given in Figure 4.8 and Figure 4.9, respectively. As can be seen in Figure 4.8, the maps of Field 1 somewhat resemble observations between 60 and 120 mg/kg, but fail to represent the maximum value between 120-180 mg/kg, and only the OK map represents the values in the lowest range of 40-60 mg/kg. The poor representation of low and high values further highlights the smoothing effect of these prediction methods. The distribution patterns of Cubist and RF are quite like each other, meaning that the values used during training might not have been particularly descriptive. In Field 2 (Figure 4.9), the OK and RF maps do not have pixels representing the lowest values (0-40 mg/kg), whilst the Cubist map created pixels with a value higher (120-180 mg/kg) than what existed in the observations. The ML DSM maps showed much more spatial variation than the OK map in Field 2.

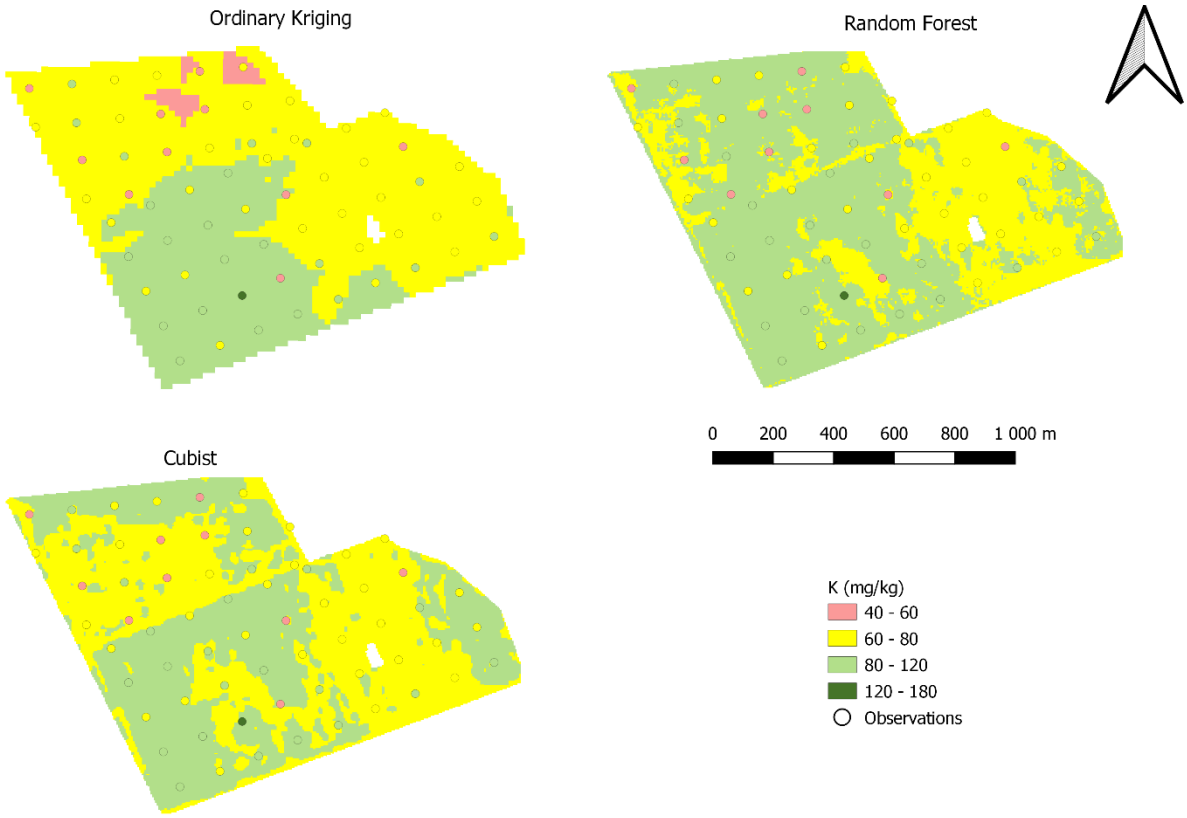

Figure 4.8: OK and ML DSM maps created of K for Field 1.

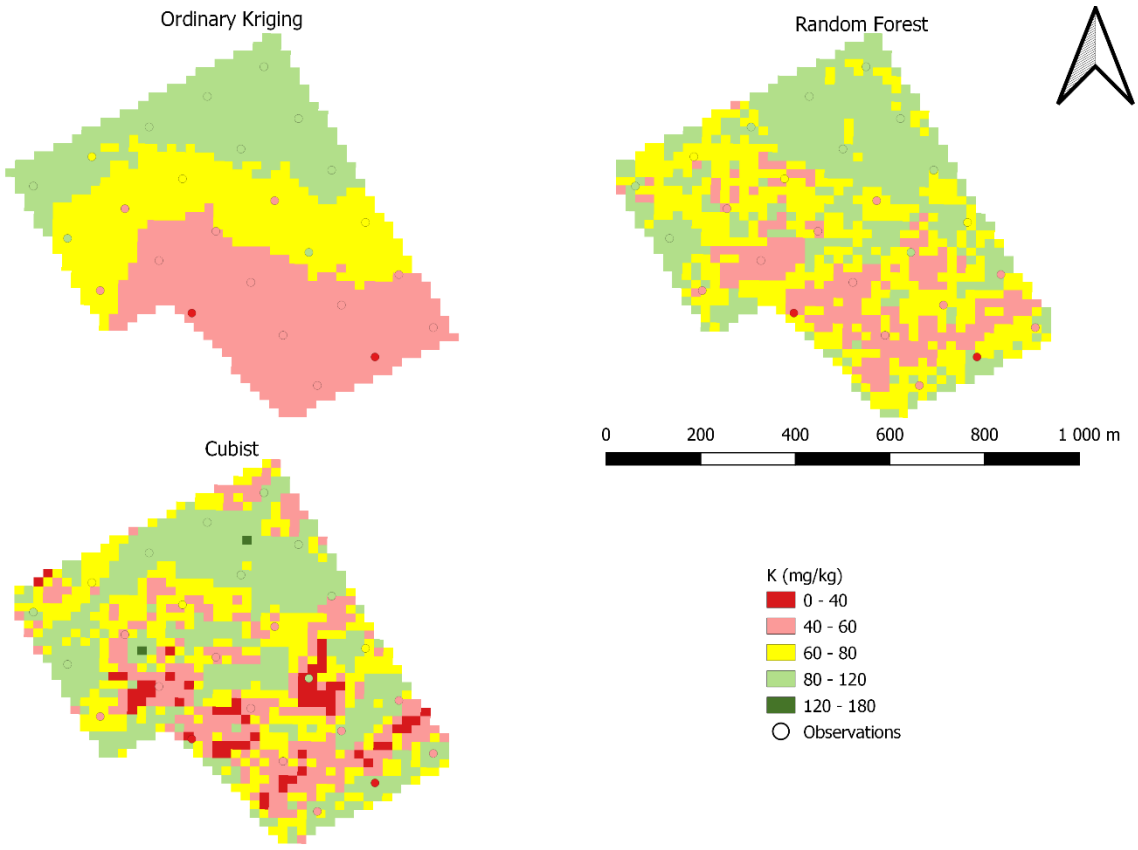

Figure 4.9: OK and ML DSM maps created of K for Field 2.

4.2.4 Ca

In Field 1, poor validation results were achieved by OK with an R^2 of 0.16 and a CCC of 0.25 (Table 4.1), while in contrast, moderate validation results were achieved in the case of Field 2 with an R^2 of 0.66 and a CCC of 0.52 (Table 4.2). The poor R^2 and CCC results of Field 1 imply weak ability to explain variation within the field and a low correlation between observation and predictions by OK, while the moderate validation results of OK in Field 2 imply a moderate explanation of variability by the model as well as a moderate agreement between observations and predictions (Khaledian & Miller, 2020). Because of these validation results, OK is not accurate enough for use in PA in Field 1, while the moderate validation results of OK in Field 2 are good enough for use in PA. The RMSE of OK in Field 2 is 75.21 mg/kg, which is around the mean of the observations at 262.63 is relatively high. As the property class used by FERTASA (2007) at this value is 250-350 mg/kg Table 3.4, the prediction error is in the same order of magnitude as the mapping class, and it is possible that the wrong fertiliser recommendation could result from the OK map.

Both Cubist and RF had poor R^2 and CCC validation results in Field 1 (Table 4.1). Cubist has an R^2 of 0.06 and a CCC of -0.18, whilst RF has an R^2 of 0.16 and a CCC of -0.25. In both cases, the implication is that variation is poorly captured and that a poor, negative correlation exists between observations and predictions. The accuracy of neither model is sufficient for use in PA alongside OK. In Field 2, Cubist was the best performing ML DSM technique with a moderate R^2 of 0.44 and a moderate CCC of 0.50 as compared to the poor R^2 of 0.10 and CCC of 0.12 achieved by RF (Table 4.2). The RMSE of the cubist model in Field 2 is 55.94 mg/kg, which is still relatively high around the previously mentioned mean of Ca for Field 2, though lower than that of OK. In Field 2, OK achieved slightly better validation results than the Cubist model in terms of R^2 and CCC, but had a slightly higher RMSE.

The OK and ML DSM prediction maps of Ca in Field 1 and Field 2 are given below in Figures 4.10 and 4.11. The smoothing effect of these prediction techniques mentioned in 4.1.2 is again visible in all maps in Figure 4.10 and for the ML DSM technique maps in Figure 4.11, with both the highest and lowest observation values not represented in the maps. In Field 1, each map is unique, with little similarity between them. In Field 2 (Figure 4.11), the OK map very closely follows the spatial distribution of Ca observations compared to the RF and Cubist maps. This further shows OK to be the better prediction method for Ca in Field 2.

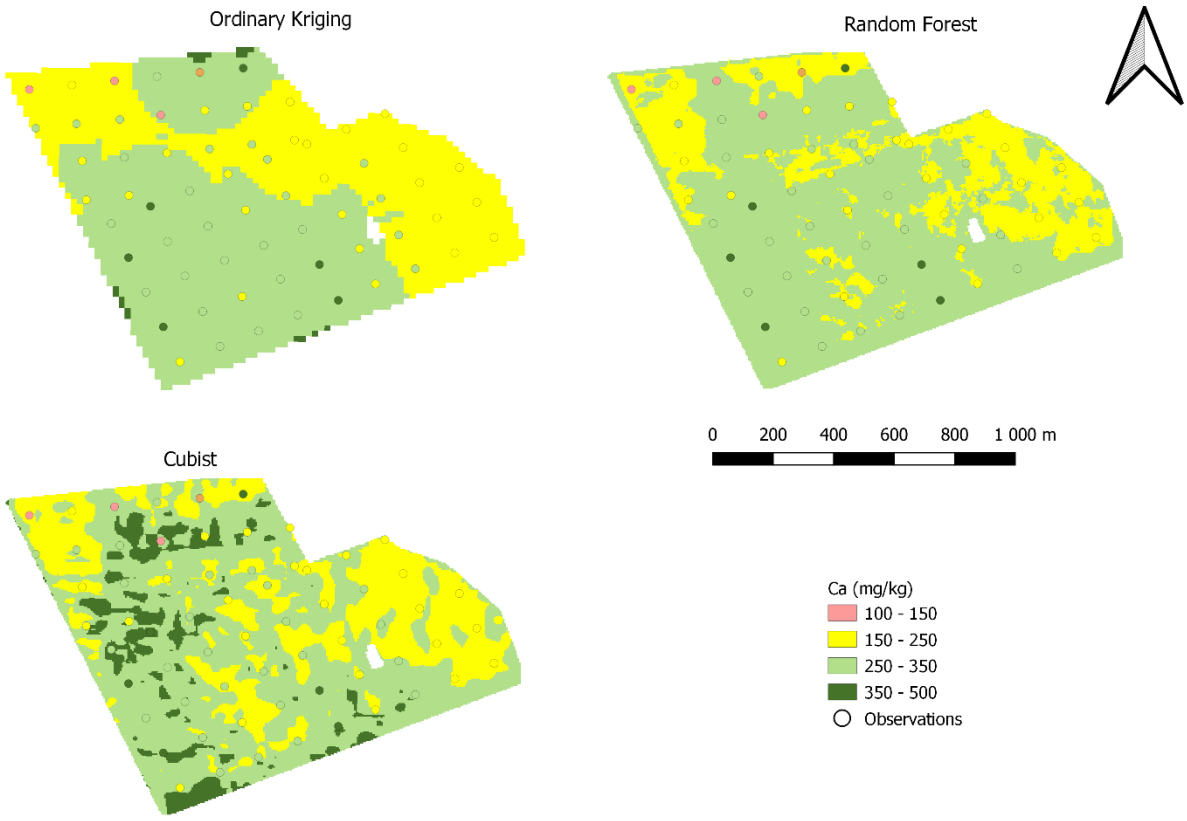

Figure 4.10: OK and ML DSM maps created of Ca for Field 1.

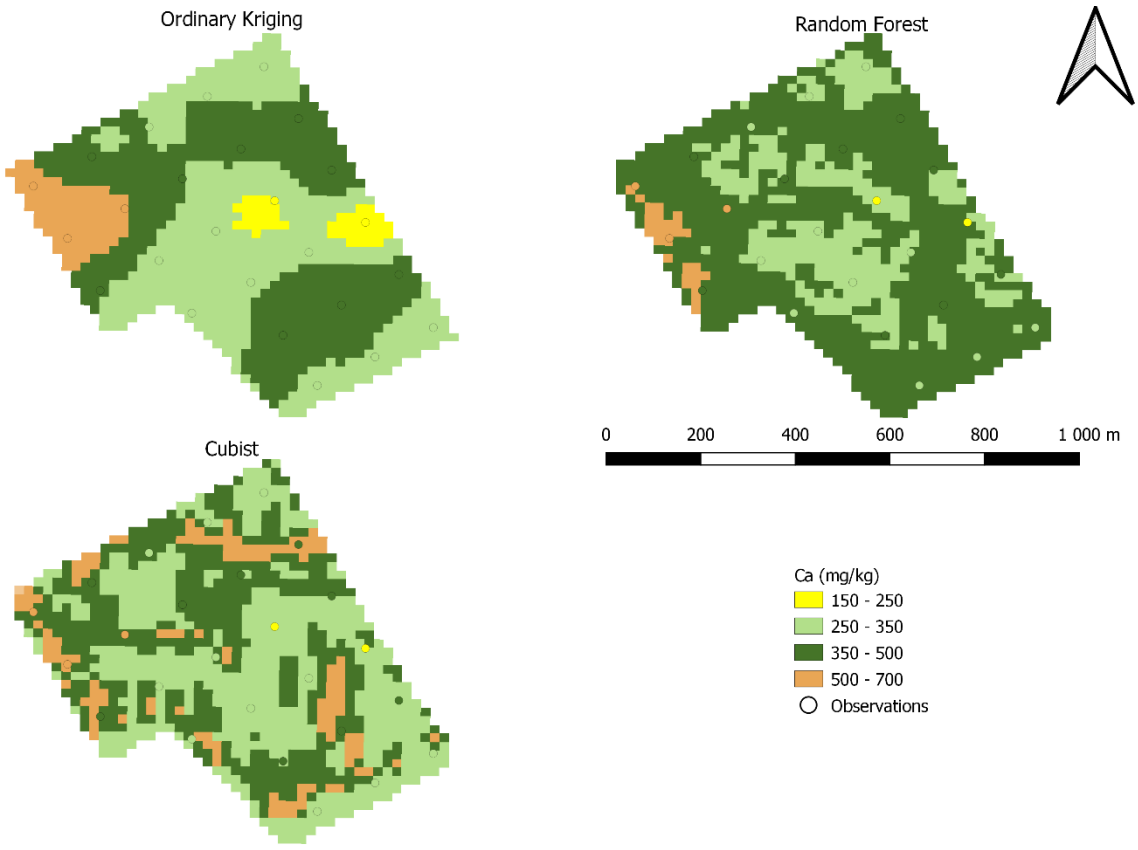

Figure 4.11: OK and ML DSM maps created of Ca for Field 2.

4.2.5 Mg

In Field 1, a moderate R^2 of 0.49 was achieved alongside a poor CCC of 0.38 (Table 4.1). Thus, although a moderate amount of variation is explained by the OK model, there still exists a poor correlation between observations and predictions (Khaledian & Miller, 2020). The RMSE of 14.73 mg/kg around the mean of the observations (83.64 mg/kg) is in the same order of magnitude as the property map class used by FERTASA (2007), between 60-80 mg/kg and 80-150 mg/kg. It is thus possible for wrongful fertiliser recommendations to be made at this point. When the relatively high RMSE and poor CCC are considered alongside the moderate R^2, OK's accuracy for use in PA is doubtful. OK in Field 2 achieved better validation results with an R^2 of 0.63 and CCC of 0.57 (Table 4.2). These results imply a good explanation of variation by OK and a moderate agreement between observations and predictions. (Khaledian & Miller, 2020). The RMSE (15.64 mg/kg) is quite low when considered alongside the mean of the observations (58.20 mg/kg) at the mapping class used by FERTASA (2007). Because of this alongside it's accuracy, the OK map of Mg would be accurate enough for use in PA in Field 2.

Both Cubist (R^2: 0.03 and CCC: -0.09) and RF (R^2: 0.0 and CCC: -0.01) performed poorly in terms of validation statistics in Field 1 (Table 4.1), and neither method could be considered accurate enough for use in PA. In Field 2, RF was the best-performing ML DSM model with a poor R^2 of 0.26 and poor CCC of 0.20 (Table 4.2), while the Cubist model failed to create a model for predicting Mg, as can be seen in Figure 4.3e, where the only value is the mean. The poor R^2 and CCC of the RF model imply a low explanation of variability by the model, as well as a poor agreement between observations and predictions, and as such, the prediction map isn't suitable for use in PA (Khaledian & Miller, 2020).

OK outperformed RF in Field 1 in terms of validation results, but as can be seen in Figure 4.2, both models failed to represent the low and high values of Mg shown in the observations. OK greatly outperformed RF in Field 2, although the same problem of not representing the lowest Mg values in the prediction persisted, hinting at the problems of data smoothing in these prediction models.

The prediction maps of Mg for Fields 1 and 2 are given in Figures 4.12 and 4.13, respectively. The smoothing effect is displayed by the lowest observations in Field 1 (20-40 mg/kg) not being represented in the prediction maps (Figure 4.12); there are also no major similarities in the distribution patterns between the maps. Little similarities between the observed Mg and prediction maps are visible for Field 1. In the cubist map of Field 2 (Figure 4.13), the prediction is of a uniform value. OK and RF followed a similar pattern in Field 2, although RF showed a lot more spatial variation. The lowest values of 40-60 mg/kg Mg in the observations are not represented in the prediction maps.

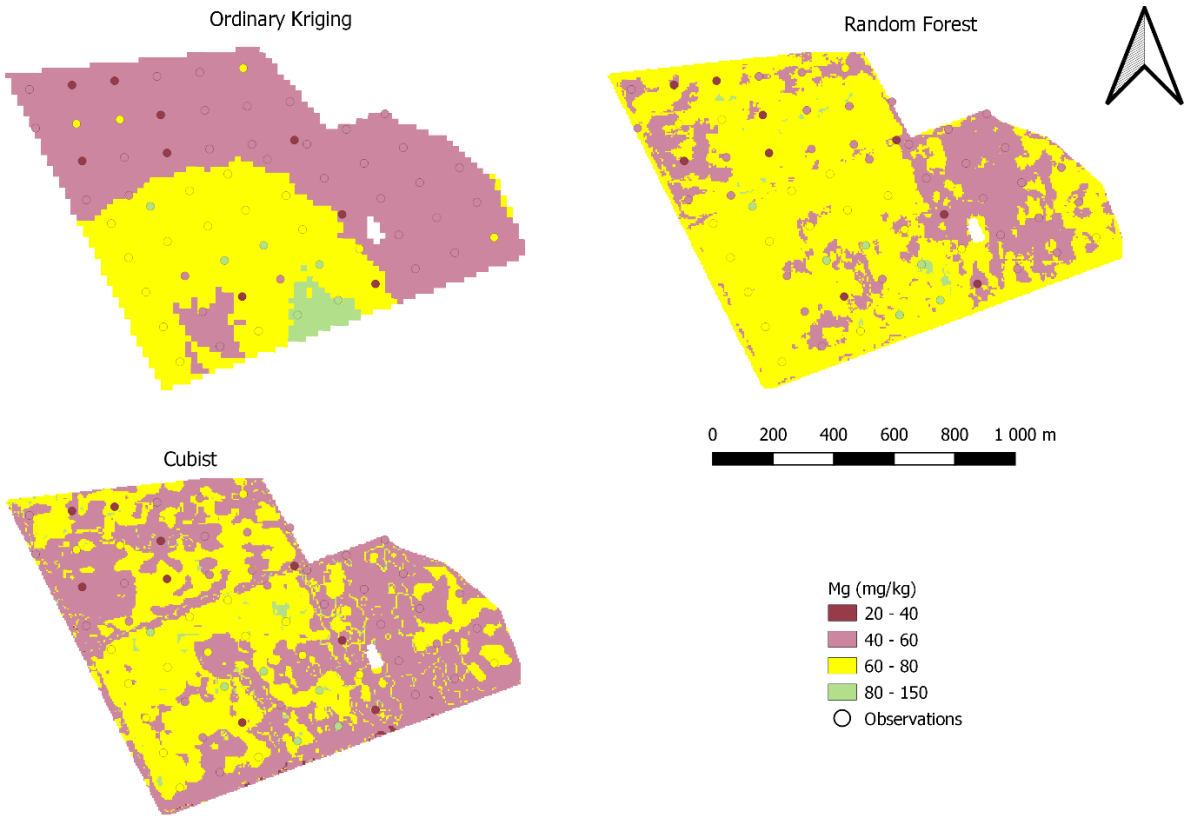

Figure 4.12: OK and ML DSM maps created of Mg for Field 1.

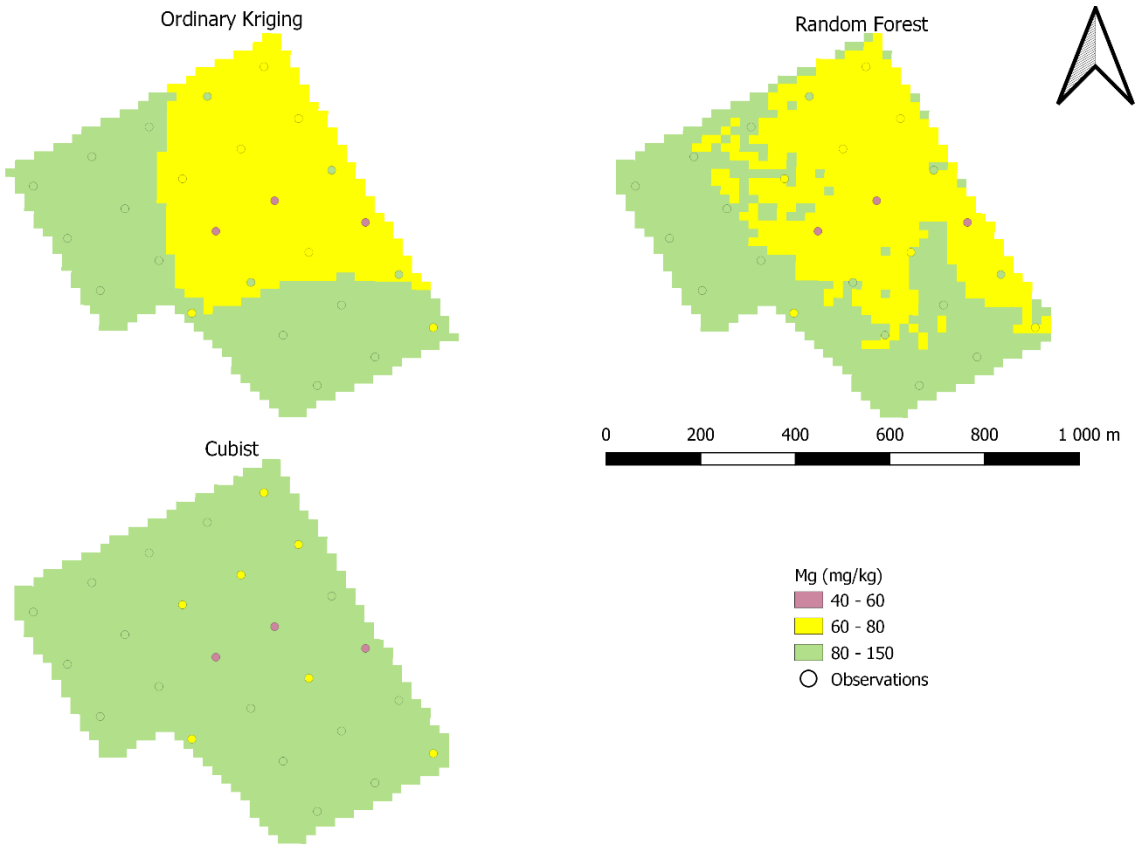

Figure 4.13: OK and ML DSM maps created of Mg for Field 2.

4.2.6 Na

Very poor validation statistics were achieved for OK in Field 1 with an R^2 of -0.3 and CCC of -0.01 (Table 4.1), meaning that very little variation was explained by the model and that there was poor agreement between observations and predictions of Na (Khaledian & Miller, 2020). The OK prediction of Na is thus not usable in PA. It should be noted that the observed Na values in Field 1 (Figure 4.2f) and Field 2 (Figure 4.3f) are, however, all within the optimal range of Na according to FERTASA (2017). Better validation results were achieved in Field 2 (Table 4.2) with a moderate R^2 of 0.48 and a poor CCC of 0.33. Implying that although some variance was accurately captured, the agreement between observations and predictions was poor, and the model is generally unreliable. The RMSE is also very low in both cases (Field 1: 0.79 and Field 2: 0.86).

The Cubist and RF models in Field 1 showed equally poor validation results with Cubist having a poor R^2 of 0.02 and poor CCC of -0.11 whilst Rf has a poor R^2 of 0.03 and poor CCC of -0.04 (Table 4.1). In the case of Field 2 (Table 4.2), RF achieved moderate to good performance with

an excellent R^2 of 0.81 and a moderate CCC of 0.66. Cubist only achieved a moderate R^2 of 0.50 and poor CCC of 0.46.

The prediction maps of Na in Field 1 are given in Figure 4.14 and of Field 1 in Figure 4.15. In Field 1, there isn't a singular definable pattern between the ML DSM techniques, with the Cubist map showing more variation than RF and OK. OK, and the ML DSM maps in Field 2 (Figure 4.15) have a similar distribution, though the distribution pattern between Cubist and RF is more alike.

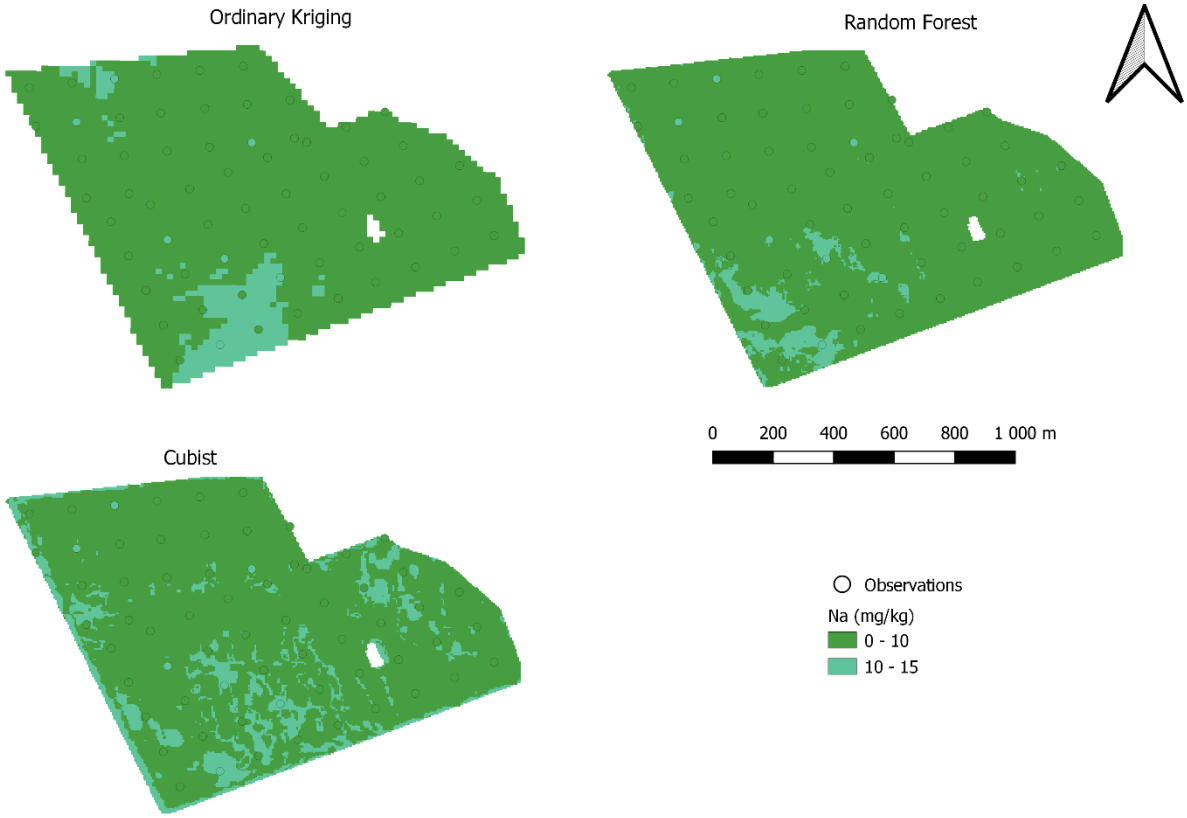

Figure 4.14: OK and ML DSM maps created of Na for Field 1.

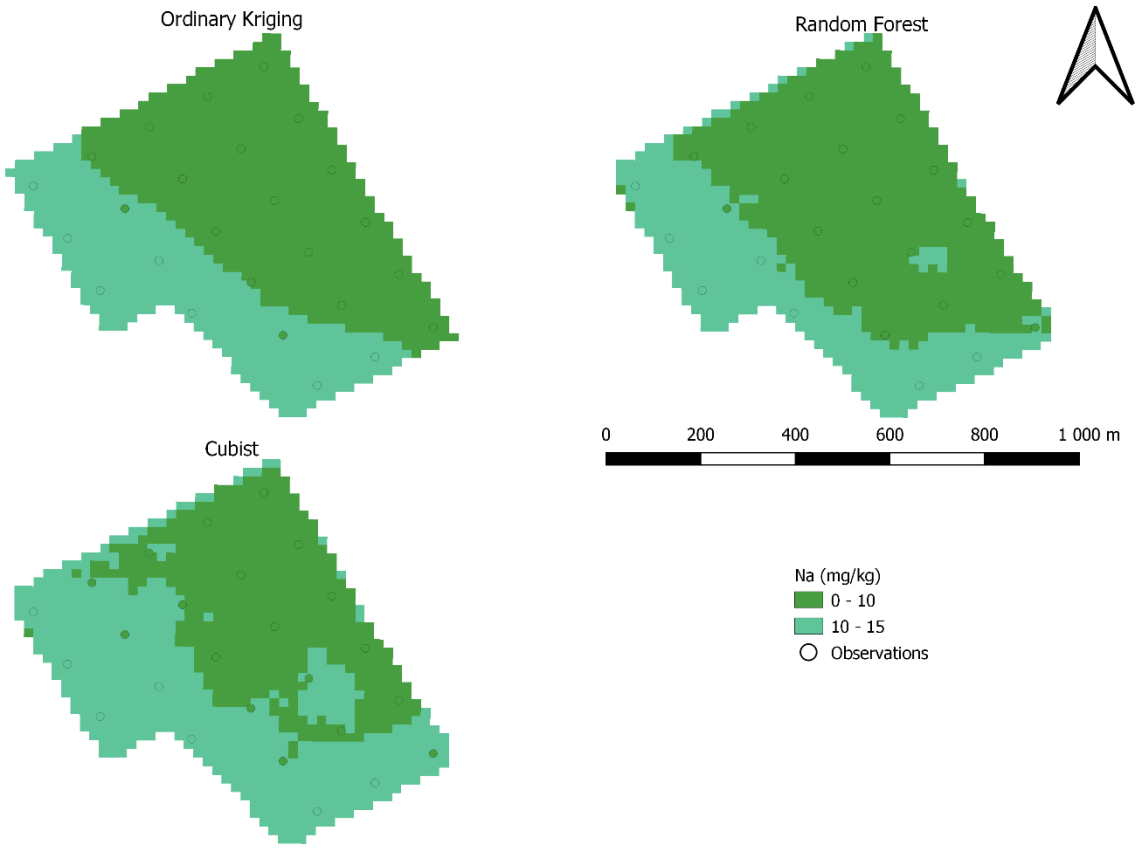

Figure 4.15: OK and ML DSM maps created of Na for Field 2.

4.2.7 OK and ML DSM

In Field 1, only the Mg OK map was accurate enough for use in PA, whilst in Field 2, the acceptable maps for OK were pH, K, Ca and Mg. This means that the accuracy of OK at the 2-ha grid used differed between fields for each element. In general, the OK maps weren't accurate enough for use in PA. This is likely because the 2-ha grid is too large for the variograms of the Ok model to capture the spatially dependent distance variation of each element. The variograms are discussed more completely in the following section. It thus becomes important to investigate the smaller grid sizes. The poor validation results of K and P in Field 1 coincide with the findings of Mallarino and Wittry (1997), that K and P levels are no longer accurately described at sampling distances greater than 0.8 ha. Another important factor that could have influenced the OK accuracy is the sampling scheme (Pereira *et al.*, 2013), meaning the grid pattern might not have been optimal. The ML DSM techniques only outperformed OK with sufficient accuracy for use in PA in the case of Na in Field 2. The only instances of ML DSM techniques showing potential for use in PA were in Field 2 with Cubist for Mg and then both Cubist and RF for Na. As the covariates used for predictions include the most important covariates named by McBratney *et al.* (2003); Khaledian and Miller (2020) and Pusch *et al.* (2022), the failure of these methods are likely

because of a weak correlation between the soil chemical properties and covariates stemming from the grid pattern which isn't optimized for using ML DSM techniques. ML DSM techniques tend to perform better when sampling is based on novel sampling designs such as those used by Yang *et al.* (2022) and Saurette *et al.* (2024). In both aforementioned cases, covariate data is used alongside algorithms like, but not limited to, Conditioned Latin Hypercube sampling (Minansny and McBratnet, 2006), to optimise the sampling location of data points for use in the training set of the ML DSM method. At this point, it should again be mentioned that the comparison between OK and the ML DSM prediction techniques in this study isn't a one-to-one comparison, as different validation techniques were used due to data availability. The obtained results do, however, indicate that Cubist and RF are not better at predicting soil property variance within fields at the farm scale in South Africa using legacy data, compared to OK.

4.3 Determining the optimal grid spacing for precision agriculture soil property mapping

Semivariograms for each of the mapped properties for each field (Figures 4.16 and 4.17; Table 4.3) have the potential to show what the optimal grid spacing is for soil property mapping in PA. According to Flatman and Yfantis (1984), the optimal sampling interval is equal to half of the range. First, the difference between the theoretical and empirical semi-variograms needs to be discussed. The empirical semi-variogram, represented by the data points on the semi-variogram figure, is the calculated representation of spatial autocorrelation of the observed data. The theoretical semi-variogram, represented by the fitted line through the data points, is a model fitted to the empirical semivariogram (Whelan & Taylor, 2013). This model is continuous and is then used during interpolation. As a shape for the theoretical variogram must be assigned, it can happen that there is a large difference between the theoretical and empirical semi-variograms (Marchant & Lark, 2004), which indicates that there is no spatial correlation for the soil property involved in that field (Marzban & Sandgathe, 2009).

As can be seen in Figure 4.16, the empirical and theoretical semivariograms of P and Mg follow each other well, whilst K, pH, Ca and Na show a clear difference between the two. In Field 2 (Figure 4.17), the empirical and theoretical semivariograms that correspond are those of K, pH, Ca and Mg, whilst the only Na does not. In the cases in Field 1 and Field 2 where the empirical and theoretical variograms did not correspond, the validation results were all poor, except for P in Field 1, which still achieved only poor validation results. Mg in Field 1, along with the other properties in Field 1, previously named, all achieved moderate validation results.

In the cases where the empirical and theoretical semivariograms correspond, the semivariogram ranges (Table 4.3) can be used to determine the optimal grid sizes for sampling density based on Whelan and Taylor's (2013) recommendation. In the case of pH in Field 2, the half range is 84.47 meters, which is much shorter than the average distance between samples on a 2-ha grid, which is 141.42 meters. When interpolating with soil data on a one-hectare grid Bogunovic *et al.* (2017) found the range for pH to be 595 meters (half range: 297.5 m), further showing the diverse nature of soil attributes. P in Field 1 has a half range of 65.09 meters, again much shorter than the average distance on a 2 ha grid, which is similar to the half of the range reported by Guo-Shun *et al.* (2010) at 62.45 meters for available P. The half range of K in Field 2 was quite large at 6852.07 meters, which could mean that little variation of K was present in the Field. On a 20 by 20 meter grid Guo-Shun *et al.* (2010) reported a half range of 156.2 meters. For Ca in Field 2, the half range was 69.56 meters and Mg in Field 1 and 2 had similar half ranges of 188.21 meters and 178.84 meters, respectively. Based on these results, it is clear that the recommended range between soil properties differs between fields and between themselves; the optimal range should then be based on the most sensitive property, which in this case was available P at 62.45 meters. As such, better kriging results will likely be obtained on sampling densities closer to half a hectare than the 2-ha grid spacing. This is like the recommendation by Brouder & Morgan (2000), grids no larger than 70 meters.

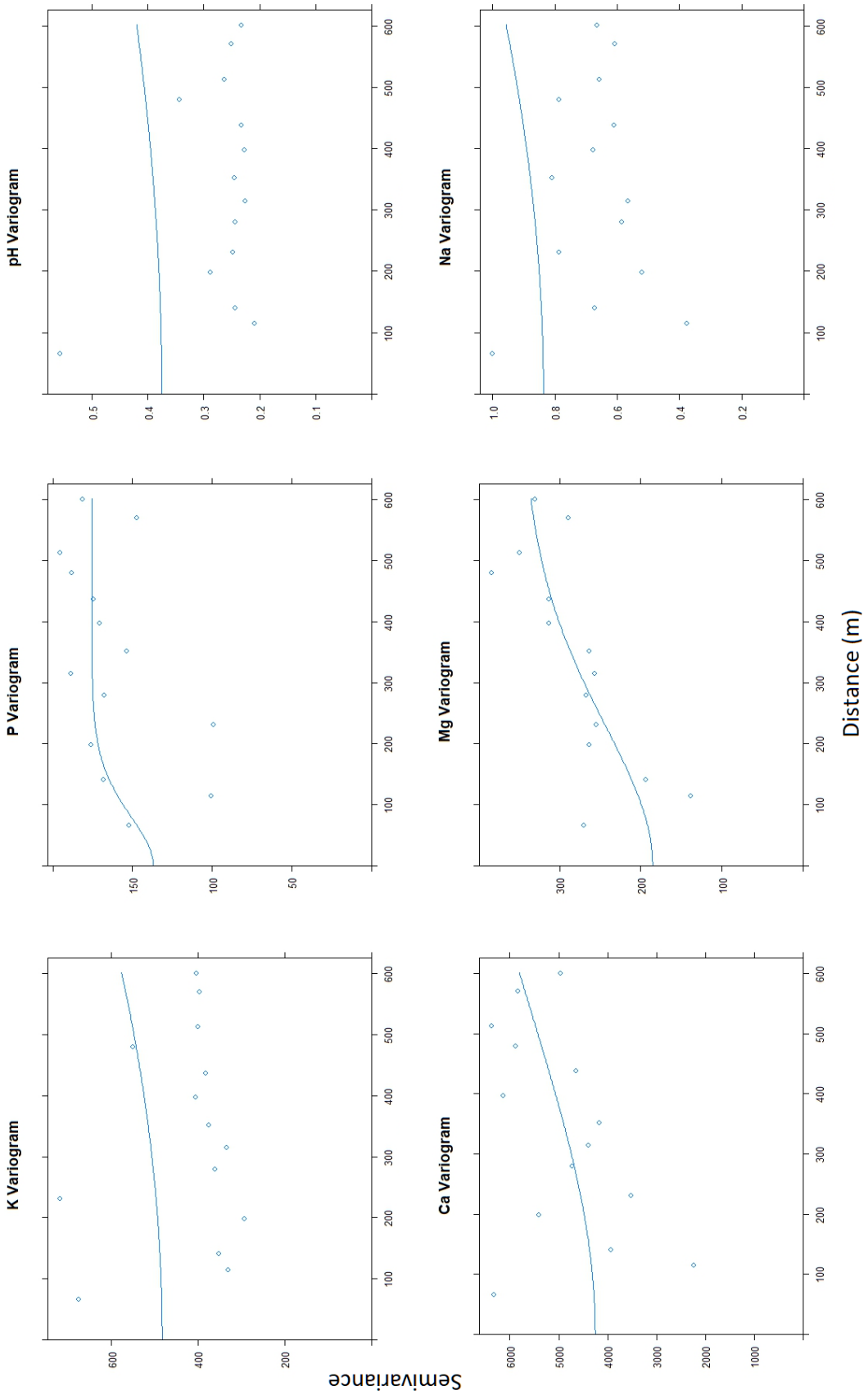

Figure 4.16: Variograms generated for the different elements in Field 1.

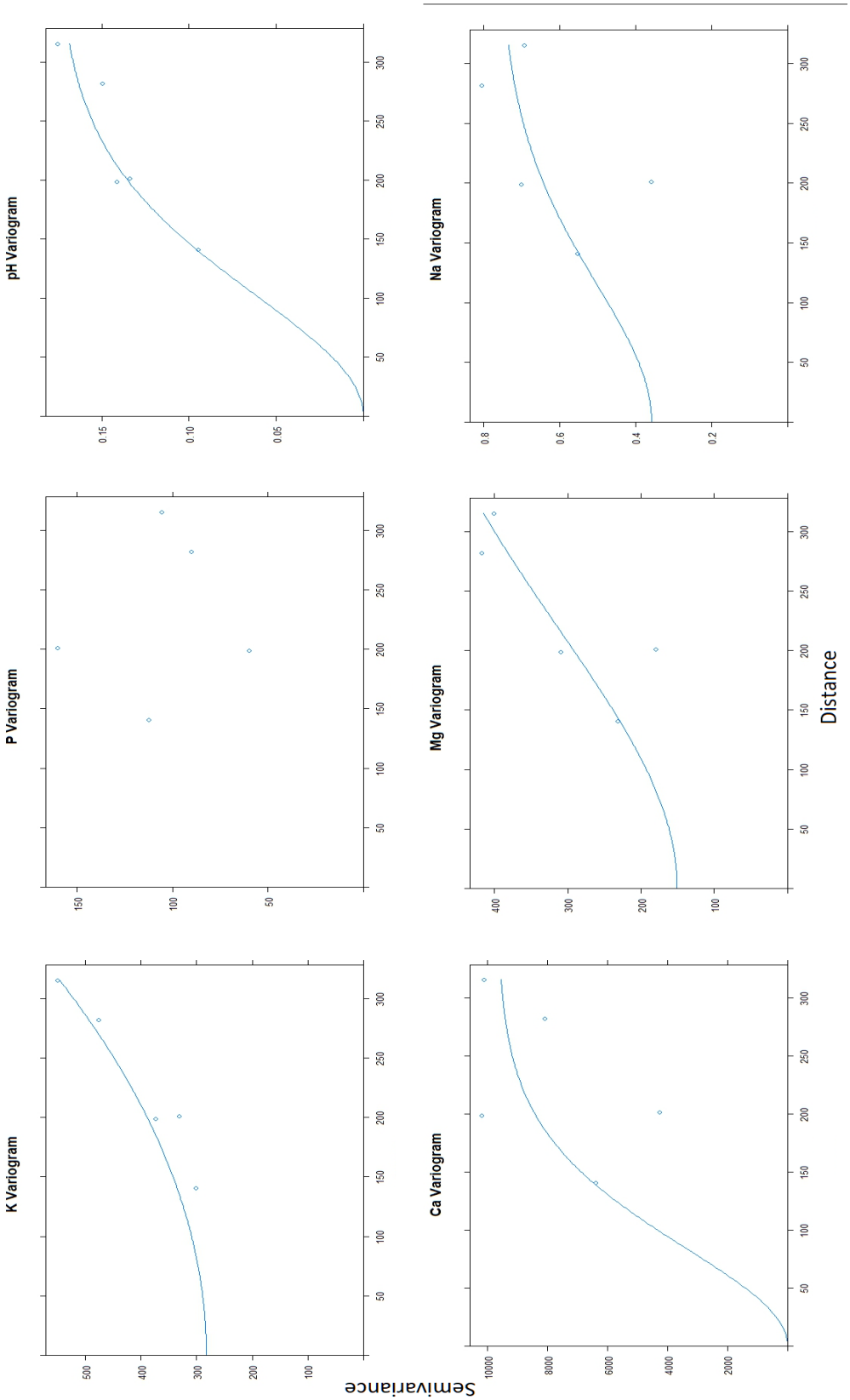

Figure 4.17: Variograms generated for the different elements in Field 2.

Table 4.3: Accompanying statistics of variograms of Fields 1 & 2

	Element	Model type	psill	Range (m)	kappa	½ Range (m)
Field 1	K	Ste	2656,51	3548,82	5,00	1774,41
	P	Ste	38,56	130,18	5,00	65,09
	pH	Ste	0,75	2746,51	5,00	1373,26
	Ca	Ste	3815,93	897,30	5,00	448,65
	Mg	Ste	164,28	376,42	5,00	188,21
	Na	Ste	3,75	4309,02	2,40	2154,51
Field 2	K	Ste	250028935,00	13704,13	2,00	6852,07
	P	Ste	72,04	813,42	5,00	406,71
	pH	Ste	0,18	168,93	5,00	84,47
	Ca	Ste	9682,46	139,12	5,00	69,56
	Mg	Ste	456,23	357,69	5,00	178,84
	Na	Ste	0,41	188,49	5,00	94,25

Ste = Stein's parameterisation model

CHAPTER 5 CONCLUSION

Three objectives were set out at the start of the study. The first objective was to test whether the conventional industry standard OK method on a 2-ha grid could produce accurate enough maps for use in PA. To test this, OK maps were generated for several soil chemical properties over two fields using results from soil samples taken on a 2-ha grid. The accuracy of these maps was then inspected through leave-one-out cross-validation, visual inspection and through analysing the semivariograms of each model. Through this inspection, it was found that OK accuracy varied between fields and soil properties. Most of the OK maps were not accurate enough for use in PA, based on poor R^2 and CCC results. The implication of this is that the current industry reliance on 2 ha and 1 ha soil property grids for use in PA in South Africa would need to be reevaluated, as these maps might not be as accurate as believed.

The second objective was to determine the optimal grid spacing for soil property mapping using OK for PA. This was done by investigating recommendations in the literature, and it was found that an optimal grid size would be half of the range of the semivariogram. So, the half range of each effective semivariogram was investigated, and it was found that the 2-ha grid pattern was likely too large, and that a 70 by 70-meter grid would probably give the most accurate results in these two fields. It was also found that the optimal range varied between soil properties and fields.

The final objective was to determine the potential for the use of ML DSM for mapping soil properties in PA. To determine this, two ML DSM techniques, namely Cubist and RF, were used to generate prediction maps for soil properties using an assortment of environmental covariates. The accuracy of these maps was then determined by splitting the dataset into training and validation sets at a 70/30 split. The accuracy of these maps was then also compared to the OK maps, alongside a visual inspection of the maps. It was found that in this case, neither model could produce accurate enough prediction maps for use in PA based on the legacy data available for the study.

Several limitations became apparent as the study continued. Firstly, the use of legacy soil property data impeded the investigation into the accuracy of OK and its comparison to the ML DSM methods. This is because two different validation measures had to be used as the existing data set already bordered on the acceptable inter-sample range. So, observations could not be removed without seriously impacting OK, to form a single validation set for comparison. As some time has already passed between the sampling dates and this study, an independent sample set could not be taken. Another possible hindrance is small-scale variance, hinted at by the variance in yield and soil type shown in the covariate data used for the ML DSM methods. As both fields consisted of various soil types, with different compositions, the 2-ha grid could not capture the resulting variance in soil properties. In the case of the ML DSM methods, it is quite possible that the models suffered because the sampling density did not sufficiently serve to capture the relation between covariates and the soil properties. This is exacerbated when not enough samples are available to capture local variation within the different areas of the field. When this happens, the ML DSM methods will struggle to find clear relationships between predictors and soil properties. The Fuzzy-K means clustering method also might have negatively influenced the accuracy of these methods. As many of the lowest and highest values weren't represented in the prediction maps, it is possible that these values were removed from the training set during data partitioning. Another limitation comes from the ML methods used. Neither Cubist nor RF allow for the setting of variable importance, and possible strong predictors might go underutilised. Lastly, the data set consisted of relatively few data points that would hamper the performance of the data quantity-sensitive ML DSM methods.

Possible future studies could attempt to test the accuracy of OK on denser grids or with different sampling methods than the traditional grid-based method. An exploratory data analysis could also be used to first determine areas in a field with high soil property variability and then increase the sample density in these specific areas. Other kriging methods, like regression- or co-kriging, could also be explored as these methods can make use of environmental covariates to improve prediction accuracy. In the case of the ML DSM methods, future studies should make use of

specialised algorithms that make use of environmental covariates or expert knowledge-based systems to determine sample locations. Finding several fields nearby could also be used to increase data quantity without necessarily increasing sampling density. Determining important covariates and reducing weak predictors could also help to further improve prediction accuracy. Different ML DSM techniques, like multinomial logistic regression or even more specialised RF models, should also be considered. To properly compare the different methods, an independent dataset should also be used. As cost constraints play a significant role in agriculture, the effect of combining sampling between different close farms and the effect of input-output ratios should also be explored. Further exploration of different landscapes and crop production systems should also be considered.

This study served to highlight the disparity between what is conventionally considered accurate enough for use in PA and the actual accuracy of the predictions made. In this study, it is clearly shown that the 2-ha soil sampling grid does not provide accurate enough OK prediction maps for use in PA. The variance of soil properties varied between fields and each other, implying that no one-size-fits-all solution for soil property predictions exists. It is also clear that simply using legacy soil data does not provide constructive data for ML DSM techniques. When attempting to create these ML DSM prediction maps, sampling methods tailored to it should be used. Further investigations into soil property mapping techniques will prove paramount to PA in South Africa's private agricultural service industry.

BIBLIOGRAPHY

Benedikt Gräler, Edzer Pebesma and Gerard Heuvelink, 2016. Spatio-Temporal Interpolation using gstat. The R Journal 8(1), 204-218

Bobryk, C.W., Myers, D.B., Kitchen, N.R., Shanahan, J.F., Sudduth, K.A., Drummond, S.T., ... Raboteaux, N.N.G. 2016. Validating a digital soil map with corn yield data for precision agriculture decision support. Agronomy Journal, 108(3):957-965.

Boettinger, J., Ramsey, R., Bodily, J., Cole, N., Kienast-Brown, S., Nield, S., ... Stum, A. 2008. Landsat spectral data for digital soil mapping. Digital soil mapping with limited data:193-202.

Bogunovic, I., Pereira, P. & Brevik, E.C. 2017. Spatial distribution of soil chemical properties in an organic farm in Croatia. Science of the total environment, 584:535-545.

Bongiovanni, R. & Lowenberg-DeBoer, J. 2004. Precision agriculture and sustainability. Precision agriculture, 5:359-387.

Brouder, S. & Morgan, M. 2000. Soil Sampling and Analysis. In: Lowenberg-DeBoer, J. Precision Farming Profitability. Agricultural Research Programs, Purdue University, West Lafayette, IN, USA, 75-81.

Burlacu, G., Costa, R., Sarraipa, J., Jardim-Golcalves, R. & Popescu, D. 2014. A conceptual model of farm management information system for decision support. In. Technological Innovation for Collective Awareness Systems: 5th IFIP WG 5.5/SOCOLNET Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2014, Costa de Caparica, Portugal, April 7-9, 2014. Proceedings 5. Springer. pp. 47-54.

Cebeci, Z. 2018. Comparison of internal validity indices for fuzzy clustering. Journal of Agricultural Informatics, 10(2):1-14 doi: 10.17700/jai.2019.10.2.537

Cisternas, I., Velásquez, I., Caro, A. & Rodríguez, A. 2020. Systematic literature review of implementations of precision agriculture. Computers and Electronics in Agriculture, 176:105626.

Council for Geoscience, 2007. Geological data 1:250 000. Council for Geoscience, Pretoria, South Africa.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.3.2, Geosci. Model Dev., 8, 1991-2007, doi:10.5194/gmd-8-1991-2015. Availible at: https://saga-gis.sourceforge.io/en/index.html Date of access: 19 Nov. 2024

Dharumarajan, S., Hegde, R. & Singh, S. 2017. Spatial prediction of major soil properties using Random Forest techniques-A case study in semi-arid tropics of South India. Geoderma Regional, 10:154-162.

Dong, W., Wu, T., Sun, Y. & Luo, J. 2018. Digital mapping of soil available phosphorus supported by AI technology for precision agriculture. In. 2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics). IEEE. pp. 1-5

Dong, W., Wu, T., Luo, J., Sun, Y. & Xia, L. 2019. Land parcel-based digital soil mapping of soil nutrient properties in an alluvial-diluvia plain agricultural area in China. Geoderma, 340:234-248.

Du plessis, M. 2018. Maak grondmonster neming metode saak? Graslander Uitgawe 5: Verandering. Available at: https://viewer.joomag.com/graslander-uitgawe-5-verandering/0261287001539762375. Date of access: 20 Mar. 25

Flatman, G.T. & Yfantis, A.A. 1984. Geostatistical strategy for soil sampling: the survey and the census. Environmental monitoring and assessment, 4:335-349.

Flynn, T., Rozanov, A., Ellis, F., de Clercq, W. & Clarke, C. 2022. Farm-scale digital soil mapping of soil classes in South Africa. South African Journal of Plant and Soil, 39(3):175-186.

Gebbers, R. & Adamchuk, V.I. 2010. Precision agriculture and food security. Science, 327(5967):828-831.

Geerts, M. 2024. Why Tree-Based Methods Do Not Work Well with Geographic Data. Available at: https://www.dataminingapps.com/2024/07/why-tree-based-methods-do-not-work-well-with-geographic-data/ Date of access: 20 Mar. 25

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. Spatial Statistics, 1, 110–120.https://doi.org/10.1016/j.spasta.2012.03.002

Goktepe, A., Altun, S. & Sezer, A. 2005. Soil clustering by fuzzy c-means algorithm. Advances in Engineering Software, 36(10):691-698.

Goovaerts, P. (1997). Geostatistics for natural resources evaluation. Oxford University Press. https://doi.org/10.1093/oso/9780195115383.001.0001

Guo-Shun, L., Hou-Long, J., Shu-Duan, L., Xin-Zhong, W., Hong-Zhi, S., Yong-Feng, Y., ... Jian-Guo, G. 2010. Comparison of kriging interpolation precision with different soil sampling intervals for precision agriculture. Soil science, 175(8):405-415.

Hanesch, M., Scholger, R. & Dekkers, M.J. 2001. The application of fuzzy c-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites. Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy, 26(11-12):885-891.

Heil, J., Häring, V., Marschner, B. & Stumpe, B. 2019. Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with West African soils. Geoderma, 337:11-21.

Hijmans R. 2023. terra: Spatial Data Analysis_. R package version 1.7-39. Available at: https://CRAN.R-project.org/package=terra.Date of Access: 20 Nov. 2024

Hot, E. & Popović-Bugarin, V. 2015. Soil data clustering by using K-means and fuzzy K-means algorithm. In. 2015 23rd Telecommunications Forum Telfor (TELFOR). IEEE. pp. 890-893.

Huang, H., Yang, L., Zhang, L., Pu, Y., Yang, C., Wu, Q., ... Zhou, C. 2022. A review on digital mapping of soil carbon in cropland: progress, challenge, and prospect. Environmental Research Letters, 17(12):123004.

Jacobs, A., Van Tol, J. & Du Preez, C. 2018. Farmers perceptions of precision agriculture and the role of agricultural extension: a case study of crop farming in the Schweizer-Reneke region, South Africa. South African Journal of Agricultural Extension, 46(2):107-118.

Kaya, F., Keshavarzi, A., Francaviglia, R., Kaplan, G., Başayiğit, L. & Dedeoğlu, M. 2022. Assessing machine learning-based prediction under different agricultural practices for digital mapping of soil organic carbon and available phosphorus. Agriculture, 12(7):1062.

Keskin, H. & Grunwald, S. 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma, 326:22-41.

Khaledian, Y. & Miller, B.A. 2020. Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling, 81:401-418.

Knotters, M., Brus, D. & Voshaar, J.O. 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma, 67(3-4):227-246.

Kock, A.L., Ramphisa-Nghondzweni, P.D. & Van Zijl, G. 2024. Development of soil spectroscopy models for the Western Highveld region, South Africa: Why do we need local data? European Journal of Soil Science, 75(6):e70014.

Kotzé, J. & van Tol, J. 2023. Extrapolation of Digital Soil Mapping Approaches for Soil Organic Carbon Stock Predictions in an Afromontane Environment. Land, 12(3):520.

Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. Available at: https://doi.org/10.18637/jss.v028.i05 Date of Access: 20 Nov. 2024

Kuhn, M. and Quinlan R. 2023. Cubist: Rule- And Instance-Based Regression Modeling. R package version 0.4.2.1. Available at: https://CRAN.R-project.org/package=Cubist Date of Access: 20 Nov. 2024

Lagacherie, P. & McBratney, A. 2006. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. Developments in soil science, 31:3-22.

Land Type Survey Staff. 1972-2006. Land Types of South Africa: Digital Map (1: 250 000 Scale) and Soil Inventory Datasets. South Africa, PTA: Agriculture Research Council Institute for Soil, Climate and Water.

Liaw, A. and Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3), 18--22.

Linden, B., SLU, Olesen, S.E. & DIAS. 2003. Implementation of Precision Farming in Practical Agriculture. Dias report. Plant production., 100,

Loures, L., Chamizo, A., Ferreira, P., Loures, A., Castanho, R. & Panagopoulos, T. 2020. Assessing the effectiveness of precision agriculture management systems in mediterranean small farms. Sustainability, 12(9):3765.

Mallarino, A. & Wittry, D. 1997. Use of DGPS, yield monitors, soil testing, and variable rate technology to improve phosphorus and potassium management.

Malone, B. 2021. ithir: Soil data and some useful associated functions. R package version 1.0.

Marchant, B. & Lark, R. 2004. Estimating variogram uncertainty. Mathematical Geology, 36:867-898.

Marzban, C. & Sandgathe, S. 2009. Verification with variograms. Weather and forecasting, 24(4):1102-1120.

McBratney, A.B., Santos, M.M. & Minasny, B. 2003. On digital soil mapping. Geoderma, 117(1-2):3-52.

Minasny, B., & Mcbratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32, 1378–1388.

Mucina, L. & Rutherford, M. 2006. The vegetation of South Africa, Lesotho and Swaziland. Strelitzia 19.,(South African National Biodiversity Institute: Pretoria, South Africa). Memoirs of the Botanical Survey of South Africa,

Nenkam, A.M., Wadoux, A.M.-C., Minasny, B., Silatsa, F.B., Yemefack, M., Ugbaje, S.U., ... Bouslihim, Y. 2024. Applications and challenges of digital soil mapping in Africa. Geoderma, 449:117007.

NWK. 2024. Products and services. Available at: https://www.nwk.co.za/produkte-dienste/?lang=ZA#landboubestuursdienste. Date of access 19 Nov. 2024

Nyaga, J.M., Onyango, C.M., Wetterlind, J. & Söderström, M. 2021. Precision agriculture research in sub-Saharan Africa countries: A systematic map. Precision Agriculture, 22:1217-1236

Oliver, M. 1987. Geostatistics and its application to soil science. Soil use and management, 3(1):8-20.

Pahlavan-Rad, M.R., Dahmardeh, K. & Brungard, C. 2018. Predicting soil organic carbon concentrations in a low relief landscape, eastern Iran. Geoderma Regional, 15:e00195.

Paterson, G., Turner, D., Wiese, L., Van Zijl, G., Clarke, C. & Van Tol, J. 2015. Spatial soil information in South Africa: Situational analysis, limitations and challenges. South African Journal of Science, 111(5-6):1-7.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30: 683-691.

Pereira, G.W., Valente, D.S.M., de Queiroz, D.M., Santos, N.T. & Fernandes-Filho, E.I. 2022. Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. Precision Agriculture, 23(4):1189-1204.

Pusch, M., Samuel-Rosa, A., Oliveira, A.L.G., Magalhães, P.S.G. & do Amaral, L.R. 2022. Improving soil property maps for precision agriculture in the presence of outliers using covariates. Precision Agriculture, 23(5):1575-1603.

QGIS Development Team, 2024. QGIS Geographic Information System. Open Source Geospatial Foundation. Available at: http://qgis.org Date of access: 19 November 2024

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ Date of Access: 20 Nov. 2024

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D., 1973. Monitoring vegetation systems in the great plains with ERTS. Proceedings, Third ERTS Symposium, 1, 48-62.

Saurette, D.D., Heck, R.J., Gillespie, A.W., Berg, A.A. & Biswas, A. 2024. Sample Size Optimization for Digital Soil Mapping: An Empirical Example. Land, 13(3):365.

Schulze, R.E., Maharaj, M., Warburton, M., Gers, C., Horan, M., Kunz, R. & Clark, D. 2007. South African atlas of climatology and agrohydrology. Water Research Commission, Pretoria, RSA, WRC Report, 1489(1):06.

Scull, P., Franklin, J., Chadwick, O.A. & McArthur, D. 2003. Predictive soil mapping: a review. Progress in Physical geography, 27(2):171-197.

Sentinel Hub. 2024. EO Browser. Sentinel Hub. Available at: https://apps.sentinel-hub.com/eo-browser/ Date of access: 19 November 2024.

Soil and Plant Analysis Council, 2000. Soil analysis Handbook of reference methods. 4th ed. CRC Press LLC: Boca Raton.

Söderström, M., Sohlenius, G., Rodhe, L. & Piikki, K. 2016. Adaptation of regional digital soil mapping for precision agriculture. Precision Agriculture, 17:588-607.

Suleymanov, A., Tuktarova, I., Belan, L., Suleymanov, R., Gabbasova, I. & Araslanova, L. 2023. Spatial prediction of soil properties using random forest, k-nearest neighbors and cubist approaches in the foothills of the Ural Mountains, Russia. Modeling Earth Systems and Environment, 9(3):3461-3471.

Swanepoel, J.W.H., Swanepoel, C.J., van Graan, F.C., Alison, J.S. & Santana, J. 2018. Elementere statistiese metodes [Elementary statistical methods]. 6th ed. Potchefstroom: Ivyline Technnologies.

The Fertilizer Society of South Africa. 2007. FSSA fertilizer handbook of South Africa. 6th ed. Pretoria.

The Fertilizer Society of South Africa. 2016. FSSA fertilizer handbook of South Africa. 6th ed. Pretoria.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(sup1), 234–240. https://doi.org/10.2307/143141

Tziachris, P., Aschonitis, V., Chatzistathis, T., Papadopoulou, M. & Doukas, I.D. 2020. Comparing machine learning models and hybrid geostatistical methods using environmental and soil covariates for soil pH prediction. ISPRS International Journal of Geo-Information, 9(4):276.

USGS. 2024. Earth Explorer. United States Geological Survey. Available at: https://earthexplorer.usgs.gov/ Date of access: 19 November 2024.

Van Evert, F.K., Gaitán-Cremaschi, D., Fountas, S. & Kempenaar, C. 2017. Can precision agriculture increase the profitability and sustainability of the production of potatoes and olives? Sustainability, 9(10):1863.

Van Vuuren, J., Barnard, R. & Claassens, A. 2000. Data presentation, interpretation, and communication: Soil Sampling Under Fixed Cultivation Practices. Communications in soil science and plant analysis, 31(11-14):2055-2066.

van Zijl, G. 2019. Digital soil mapping approaches to address real world problems in southern Africa. Geoderma, 337:1301-1308.

Van Zijl, G., Van Tol, J., Bouwer, D., Lorentz, S. & Le Roux, P. 2020. Combining historical remote sensing, digital soil mapping and hydrological modelling to produce solutions for infrastructure damage in Cosmo City, South Africa. Remote Sensing, 12(3):433.

Van Zijl, G.M., Le Roux, P.A. & Turner, D.P. 2013. Disaggregation of land types using terrain analysis, expert knowledge and GIS methods. South African Journal of Plant and Soil, 30(3):123-129.

Venter, Z.S., Hawkins, H.-J., Cramer, M.D. & Mills, A.J. 2021. Mapping soil organic carbon stocks and trends with satellite-driven high resolution maps over South Africa. Science of the Total Environment, 771:145384.

Verster, E., Du Plessis, M.J. and Schoeman, J.L., 2022. Guidelines for conducting soil surveys in South Africa. Soil science society of South Africa, Bloemfontein, South Africa.

Whelan, B. & Taylor, J. 2013. Precision agriculture for grain production systems. Csiro publishing.

Xiao, S., Ou, M., Geng, Y. & Zhou, T. 2023. Mapping soil pH levels across Europe: An analysis of LUCAS topsoil data using random forest kriging (RFK). Soil Use and Management, 39(2):900-916.

Yamamoto, J.K. 2005. Correcting the smoothing effect of ordinary kriging estimates. Mathematical geology, 37:69-94.

Yang, H., Lim, H., Moon, H., Li, Q., Nam, S., Kim, J. & Choi, H.T. 2022. Simple optimal sampling algorithm to strengthen digital soil mapping using the spatial distribution of machine learning predictive uncertainty: A case study for field capacity prediction. Land, 11(11):2098.

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

Zhao, W., Ma, J., Liu, Q., Song, J., Tysklind, M., Liu, C., ... Wu, F. 2023. Comparison and application of SOFM, fuzzy c-means and k-means clustering algorithms for natural soil environment regionalization in China. Environmental Research, 216:114519.

ZHU, Q. & Lin, H. 2010. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. Pedosphere, 20(5):594-606.