



Evaluating Different Soil Mapping Methods for Precision Agriculture

J Gerald

 **orcid.org/ 0000-0002-9182-5803**

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
Co-supervisor: Dr PD Ramphisa-Nghondzweni
Assistant Supervisor: Dr WH Cloete

Graduation: November 2026

DECLARATION

I confirm that this dissertation, Evaluating Different Soil Mapping Methods for Precision Agriculture, presented for the *Master of Science in Agriculture with Soil Science* degree is entirely my own, original work and therefore has not been submitted to any other academic institution.

Furthermore, I acknowledge and agree that the North-West University possesses exclusive publication rights to this dissertation.


.....

J. Geraldès

2nd of October 2026

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the Living God, Jesus Christ, for the guidance and comfort He gave me throughout this season of my life.

Psalms 144 1(NKJV) – *“Blessed be the Lord my Rock, who trains my hands for war, and my fingers for battle- “*

I would like to thank the following people, without whom this project would not have been possible:

- Thank you to my supervisor, Professor GM van Zijl, who gave me the opportunity to achieve my MSc. Your patience is the stuff of legends and the time you took to guide me and help me through this project I am eternally grateful for.
- Dr Willie “Tsiriletso” Cloete, who took very valuable time out of his PhD work to help me along and guide me through this process. I am grateful to you wholeheartedly.
- My parents, Carlos and Barbara Geraldés, who instilled in me the work ethic that made it possible for me to complete this project, and without whose support, I would not have completed it. As well as my brother, Raphael Geraldés, for supporting me throughout.
- Gideon Nieuwoudt, for his contributions and friendship throughout this work.
- Dina Buys, who supported me without fail and with heart rarely found in this world.
- Dante Barrett, for the spiritual guidance and brotherhood.
- Dr Dimakatso Ramphisa-Nghondzweni for help with the fertiliser recommendations.
- Mr Edrich Louw for his help with programming.
- Martin Stols for supplying the soil data for this project.

ABSTRACT

This study aimed to determine the most accurate and cost-effective method of soil mapping at field scale. To achieve this, a field was selected from the Free State Province in South Africa, and soil analysis was conducted on samples taken on a 1-hectare grid basis. The industry standard method of soil mapping in South Africa, Ordinary Kriging (OK), was evaluated alongside semi-variograms to determine the viability of using 1-hectare grids to sample soils. The accuracy of the OK maps was assessed using 10-fold cross-validation. Furthermore, digital soil mapping (DSM) with Machine Learning (ML) models were used to create maps from the same properties. The two models used were Cubist and Random Forest (RF), while the covariates used included Brightness Index (BI), Redness Index (RI), Saturation Index (SI), Normalised Difference Vegetation Index (NDVI), Digital Elevation Model (DEM) and Harvester Yield Data. The accuracy of the DSM with ML maps was tested by splitting the dataset 70% - 30%, where 70% of the dataset was used to train the model and 30% was used to validate the predictions. The statistics of the OK and DSM with ML maps were compared, and the maps were also visually inspected. Thereafter, fertiliser recommendation maps were created from the potassium (K) and phosphorus (P) maps of both the OK and DSM predictions. From these fertiliser recommendation maps a financial assessment was conducted of the two mapping methods. The semi-variograms showed that the 1-hectare grid system should not be used. The potassium (K) calcium (Ca), magnesium (Mg) and cation exchange capacity (CEC) OK maps all showed directionality. Both the DSM with ML and OK results could not be considered accurate enough. The financial analysis showed that due to the maps being inaccurate, there were huge losses when using both methods of mapping.

Keywords: Digital Soil Mapping, Fertiliser Recommendations, Financial Implications, Machine Learning, Ordinary Kriging, Semi-variograms

TABLE OF CONTENTS

DECLARATION	I
ACKNOWLEDGEMENTS	II
ABSTRACT	III
CHAPTER 1 INTRODUCTION	11
1.1 Background and Problem Statement	11
1.2 Aims and Objectives	12
1.3 Hypotheses of the study	12
CHAPTER 2 LITERATURE REVIEW.....	13
2.1 Introduction	13
2.2 Precision Agriculture	13
2.2.1 Principles.....	13
2.2.2 Precision Agriculture in South Africa	13
2.3 Conventional Soil Mapping in South Africa.....	14
2.3.1 Principles.....	14
2.3.2 Advantages and Disadvantages of Conventional Soil Mapping	15
2.4 Digital Soil Mapping	16
2.4.1 Principles.....	16
2.4.2 Advantages and Disadvantages of Digital Soil Mapping	17
2.4.3 Previous Studies in Southern Africa at Field Scale	17
2.5 Machine Learning Models	18
2.5.1 Cubist	18

2.5.2	Random Forest (RF).....	18
2.6	<u>Model Validation</u>	19
2.6.1	Cross-Validation	19
2.7	Validation Statistics.....	19
2.7.1	Root Mean Square Error (RMSE)	19
2.7.2	Model Efficiency Coefficient (MEC).....	20
2.7.3	Lin's Concordance Correlation Coefficient (CCC)	20
2.8	Summary and Research Gap	20
CHAPTER 3	<u>MATERIALS AND METHODS</u>	21
3.1	Study Site and Site Description.....	21
3.2	Soil Sampling, Data Collection and Analysis	22
3.3	Evaluation of Conventional Method of Soil Mapping	22
3.4	Comparison of the Conventional Soil Mapping Method and alternative mapping techniques.....	23
3.4.1	Digital Soil Mapping	23
3.5	Quantification of Financial Implications of the Mapping Methods for Fertiliser Recommendations.....	26
3.5.1	Fertiliser Application Maps.....	26
3.5.2	Financial Simulations.....	26
CHAPTER 4	RESULTS AND DISCUSSION	29
4.1	Introduction	29
4.2	Descriptive Statistics	29
4.3	Comparison of Different Mapping Methods.....	32

4.3.1	Soil pH.....	32
4.3.2	Phosphorous	34
4.3.3	Potassium (K).....	36
4.3.4	Calcium (Ca)	39
4.3.5	Magnesium (Mg).....	40
4.3.6	Sodium (Na)	42
4.3.7	Cation Exchange Capacity.....	44
4.3.8	Summary	46
4.4	Fertiliser Recommendation and Cost-benefit Analysis	48
4.4.1	Fertiliser Recommendation Maps	48
4.4.2	Financial Implications Analysis	49
CHAPTER 5 CONCLUSIONS		51
REFERENCES.....		54

LIST OF TABLES

Table 1: Satellite Imagery Capture Dates	Error! Bookmark not defined.
Table 2: Covariates determined from satellite images (as derived from Flynn et al., 2019).....	24
Table 3: Fertiliser Price Table	28
Table 4: Descriptive Statistics of Soil Properties.....	30
Table 5: The accuracy statistics of the pH maps.....	34
Table 6: The accuracy statistics of the phosphorus (P) maps.....	36
Table 7: The accuracy statistics of the potassium (K) maps.	38
Table 8: The accuracy statistics of the calcium (Ca) maps.	40
Table 9: The accuracy statistics of the magnesium (Mg) maps.....	42
Table 10: The accuracy statistics of the sodium (Na) maps.....	44
Table 11: The accuracy statistics of the cation exchange capacity (CEC) maps.....	46
Table 13: Fertiliser Financial Implication Values	50

LIST OF FIGURES

Figure 1: Theoretical spherical semi-variogram model. C_0 = nugget, C_1 = partial sill, $C_0 + C_1$ = sill, α = range.....	15
Figure 2: Location of Study Site and Sampling Points at Study Site.	21
Figure 3: Examples of Covariate Data. NDVI, Yield and DEM Maps.....	24
Figure 4: Example of a graph used to determine the yield equations.....	28
Figure 5: Raw Data Maps.....	31
Figure 6: The constructed semi-variogram (a) and OK map (b) of pH.....	32
Figure 7: The generated pH maps of RF (a) and Cubist (b).....	34
Figure 8: The constructed semi-variogram (a) and OK map (b) of the phosphorus (P).	35
Figure 9: The generated phosphorus (P) maps of RF (a) and Cubist (b).	36
Figure 10: The constructed semi-variogram (a) and OK map (b) of the potassium (K).	37
Figure 11: The generated potassium (K) maps of RF (a) and Cubist (b).....	38
Figure 12: The constructed semi-variogram (a) and OK map (b) of the calcium (Ca).	39
Figure 13: The generated calcium (Ca) maps of RF (a) and Cubist (b).....	40
Figure 14: The constructed semi-variogram (a) and OK map (b) of the magnesium (Mg).....	41
Figure 15: The generated magnesium (Mg) maps of RF(a) and Cubist (b).....	42
Figure 16: The constructed semi-variogram (a) and OK map (b) of the sodium (Na).	43
Figure 17: The generated sodium (Na) maps of RF (a) and Cubist (b).	44
Figure 18: The constructed semi-variogram (a) and OK map (b) of the cation exchange capacity (CEC).....	45
Figure 19: The generated Cation Exchange Capacity (CEC) maps of RF (a) and Cubist (b).	46
Figure 20: Yield Potential Map.....	48

Figure 21: Map indicating the Phosphorus Fertiliser Recommendation created with conventional mapping (OK) (a) and digital soil mapping (DSM) with machine learning (ML) (b).	49
Figure 22: Maps indicating the Potassium Fertiliser Recommendation Maps created using conventional mapping (OK) (a) and digital soil mapping (DSM) with machine learning (ML) (b).	49

LIST OF EQUATIONS

Equation 1: General Spatial Model	14
Equation 2: Ordinary Kriging.....	14
Equation 3: CLORPT Model	16
Equation 4: RMSE.....	19
Equation 5: MEC.	20
Equation 6: CCC.	20
Equation 7: Extra Fertiliser Cost	27

CHAPTER 1 INTRODUCTION

1.1 Background and Problem Statement

Food security worldwide is steadily becoming a major concern, with the second goal of the United Nations Sustainable Development Goals being “Zero Hunger” (United Nations, 2023). To attain this goal, food production will need to increase drastically (Da Silva, 2012). Soil degradation, climate change and rising input prices continue to challenge farmers and food producers worldwide. Thus, it is necessary for alternative strategies to increase crop production with decreasing arable land levels. One method of solving this problem and aiding food production is through precision agriculture (PA) (Erickson & Fausti, 2021).

PA incorporates the understanding of soil, plant and field environmental conditions, which enables farmers and food producers to apply crop and soil treatments more effectively to maximise profits and yields. One major aspect of PA is the production of soil maps, which can be used for Site-Specific Crop Management (SSCM). This management strategy involves the use of soil property maps by tractors and implements with Global Positioning Systems (GPS) and application rates, which are changed real time and in-field. PA places emphasis on minimizing input wastage and avoiding over application of inputs that could damage the environment (such as chemical fertilisers) and particularly soils, which also aids in limiting the cost of inputs (Gebbers and Adamchuk, 2010).

In the South African commercial agriculture context, PA application on sampling for fertiliser recommendations. The industry standard method is to take soil samples at a one- or two-hectare grid distance (Verster et al., 2022) and then using ordinary kriging (OK) to interpolate the values between the sampling locations to obtain continuous soil property maps, for fertiliser recommendations. The two-hectare grid size was proven to be insufficiently accurate due to the distance between sampling points in a study conducted in the North West Province of South Africa (Louw, 2025), but according to my knowledge a peer reviewed independent study has not been conducted in South Africa to assess the viability of the 1 ha grid spacing. The grid sampling mapping methods and techniques have proven to be expensive, labour intensive and time-consuming (Wadoux, 2020; Nguyen, 2019; Berger, 2020). Previous studies (Mazzoncini, 2024; Lamichhane, 2019; Loiseau et al., 2021) have shown that advancing digital soil mapping (DSM) technologies incorporating machine learning (ML) techniques can be used to create maps with fewer sampling points. ML models have proved to be a powerful tool for soil mapping in general, finding use in a multitude of applications across the globe, such as mapping soil associations to inform gully erosion (du Plessis et al., 2020). Furthermore, Random Forest (RF) and Cubist were successfully used to map soil organic carbon (SOC) in a study conducted in Nigeria (John et al.,

2020). DSM found another application in mapping nine different soil properties at a far larger scale, namely, the continent of Africa. At a resolution of 250m x 250m, RF yielded acceptable results for 7 of the 9 properties (Hengl et al., 2015). Therefore, DSM with ML has been proved to be an advanced method of mapping soil properties. DSM make use of environmental covariate data (which are proxies for soil-forming factors within the CLORPT and SCORPAN frameworks (McBratney et al., 2003), such as spectral indices from satellite sensed sensors such as Normalized Difference Vegetation Index (NDVI), Brightness Index (BI) and topographic variables derived from a Digital Elevation Model (DEM) to make more accurate predictions when mapping soils (Khaledian and Miller, 2020). This can thus prove to be a major improvement in the field of soil mapping as it can provide better results at field scale than the industry standard method, thus aiding in improving food security.

To improve food security and decrease soil degradation; it is necessary to use the appropriate soil property mapping methods with an accuracy sufficient for PA grain production in South Africa at field scale. For this purpose, a field-scale case study was conducted on a 354-ha dryland grain production field located approximately 10 km south of Orkney, Free State Province, South Africa, in which OK and DSM using ML were evaluated as soil mapping methods for PA.

1.2 Aims and Objectives

This study aimed to test different soil mapping methods for PA at field scale for grain production in South Africa.

Objectives of this study included to:

- Test the industry standard method of soil mapping (OK) on a 1 ha grid to determine if it is sufficiently accurate for PA.
- Test the viability of DSM for soil property mapping for PA.
- Determine the financial implications of the two mapping methods.

1.3 Hypotheses of the study

Hypotheses of the study included:

- The industry standard method of soil mapping will not be sufficiently accurate for PA.
- DSM is a viable alternative for soil property mapping for PA.
- The mapping methods will all result in losses for the farmer as a result of inaccuracies in the mapping.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

This section will provide an overview of on the principles related to PA, DSM and conventional mapping. The advantages and disadvantages of DSM and conventional mapping will also be discussed, as well as previous studies focusing on these mapping approaches in Southern Africa.

2.2 Precision Agriculture

2.2.1 Principles

PA is an information-based approach to managing farm inputs and processes (Lowenberg-DeBoer and Erickson, 2019). It has become a crucial topic in recent years as more emphasis is placed on food security and increasing food production requirements (Grafton et al, 2015). PA is made up of several principles which have been known to reduce wastage of farm management requirements, such as fertiliser and seeds, while also increasing crop productivity (Omer et al., 2024). It has been found that smaller farms comprise the majority of all farmers in the United States of America (U.S.A.), controlling 49% of cultivated land, but only contribute 22% of all production for the U.S.A. The reason for this disparity of productivity is largely due to the reluctance of smaller farmers to adopt PA practices, whether due to a lack of knowledge or due to funding constraints (Erickson and Fausti, 2021). A drawback of PA is the reluctance of farmers to implement these new technologies due to the complexity of some of the technologies and principles, particularly in developing countries, such as South Africa (da Silveira et al., 2021). PA is split into two main subheadings, namely, crop production and livestock production (Monteiro et al., 2021). SSCM is one of the leading principles of PA with regards to crop production (Whelan & Taylor, 2010) and is based on the SSCM Cycle, which is comprised of Crop, Soil and Climate Monitoring, Attribute Mapping, Decision Support Systems and Differential Action. SSCM can be summarised as applying the right treatment, in the right amount, at the right place and the right time. Utilising these factors and principles, it is possible for farmers and agriculturalists to effectively manage crops in a cost-effective and precise manner (Erickson and Fausti, 2021). The main disadvantages of PA include the initial expense of implementing the necessary technologies, such as sensors and Geographic Information Systems (GIS). Unfortunately, GIS does also require the use of maps to effectively apply treatments in fields in the correct areas. and thus, accurate maps need to be generated for PA.

2.2.2 Precision Agriculture in South Africa

Within Southern Africa, South Africa is the country with the most application of PA, mainly on farms of a larger scale (Onyango et al., 2021). In a study conducted regarding PA adoption in the

Schweizer-Reneke region, a significant factor in the determination of whether or not a farmer introduces PA into their system is the scale of their operation, with the majority of farmers that made use of PA showed considerable cost savings as a result of PA, mainly as a result of the implementation of variable fertiliser application (Jacobs et al, 2018). The adoption of PA in South Africa is, however, increasing at a rate akin to other developing countries such as Argentina, Brazil and Turkey (Say et al., 2018).

2.3 Conventional Soil Mapping in South Africa

2.3.1 Principles

Conventional soil mapping in South Africa is conducted by sampling soils on a 1- or 2-hectare (ha) grid system (Verster et al., 2022) and utilising OK for map creation. However, a recent study by Cloete et al. (2025) showed that this sampling strategy might not effectively capture the spatial structure of soil properties, specifically at field scale. OK is a method of interpolation that considers the spatial autocorrelation described by a semi-variogram that is based on the sample points to make a prediction of the values at unsampled locations (Ryu et al., 2002). This is carried out by assigning Kriging weights to points according to a created semi-variogram model using the sampled points. When conducting predictions in geostatistics, the general spatial model for an unknown variable is defined in Equation 1, where $Z(s)$ represents the variable of interest, μ represents the (constant) mean, or trend component, and $\varepsilon(s)$ represents the autocorrelated error (Environmental Systems Research Institute, 2025).

Equation 1: General Spatial Model

$$Z(s) = \mu + \varepsilon(s)$$

This general model underlies Ordinary Kriging, the prediction formula for which is presented in Equation 2. The number of points associated with the prediction can be set using a search window, which is a radius assigned automatically or manually, depending on the specific tool in use (Webster and Oliver, 2007).

Equation 2: Ordinary Kriging

$$\hat{z}_k = \sum_{i=1}^n \lambda_i z(s_i)$$

where λ_i is the weighting coefficient for a particular location, $z(s_i)$ is the sampled value, and $\hat{z}_k$ is the prediction at unsampled locations (Webster and Oliver, 2007).

Semi-variograms are essential to the process of mapping soil properties using OK as they describe the spatial correlation structure of the data, after which an weight is applied to the points that are then used when making the predictions (Webster and Oliver, 2007). Importantly, an appropriately estimated and fitted semi-variogram model is essential for accurate OK predictions. Figure 1 displays an example of a theoretical semi-variogram model, in which the range represents the distance at which the samples are no longer spatially correlated, while the nugget represents the discontinuity at the origin of the semi-variogram, which reflects the spatial variation occurring at distances smaller than the sampling resolution, as well as potential measurement error. Under the assumption of second-order stationarity, the sill also represents the variance (Webster and Oliver, 2007).

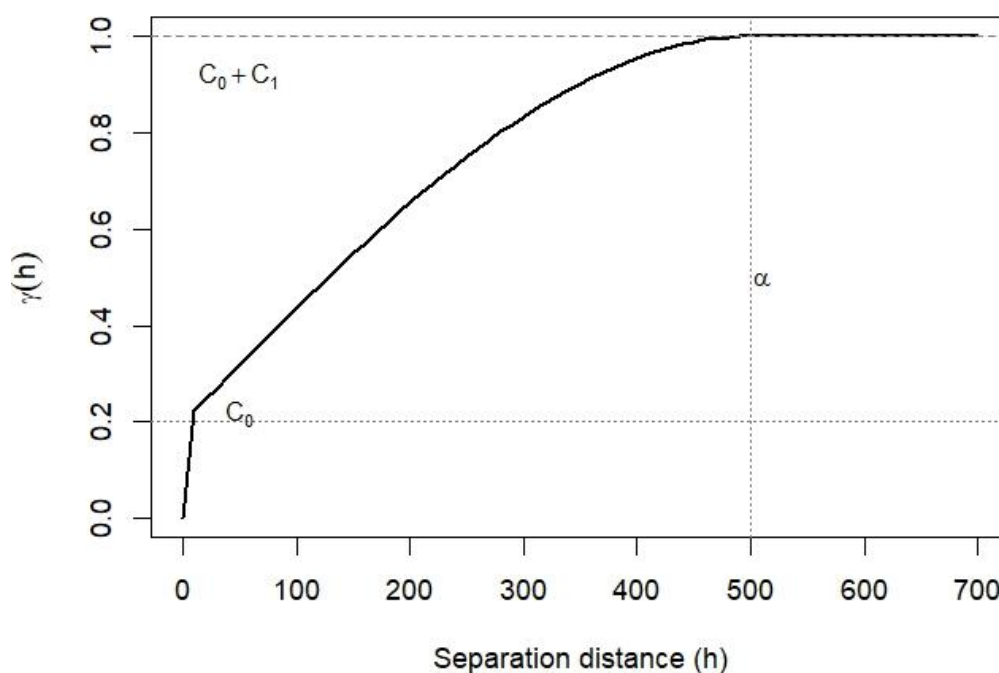


Figure 1: Theoretical spherical semi-variogram model. C_0 = nugget, C_1 = partial sill, $C_0 + C_1$ = sill, α = range.

2.3.2 Advantages and Disadvantages of Conventional Soil Mapping

One of the advantages of OK is that it is not reliant on covariate data and is capable of making predictions based on spatial autocorrelation alone. A disadvantage of OK is that it is dependent on the accurate modelling of semi-variograms. This is proving to be a challenge in South Africa's PA industry, as most companies/agriculturalists make use of readily available automated mapping softwares, which do not include post-fit examination of semi-variograms in order to determine goodness-of-fit. Additionally, OK cannot take into consideration the use of other predictor variables, such as satellite imagery, to make predictions (Meng et al., 2013). Other methods of

Kriging, such as Regression Kriging and Universal Kriging, are able to use auxiliary data, but are underutilised in South Africa in PA. This underutilisation of other Kriging methods can be attributed to the additional technical expertise, as they require more steps to conduct, such as covariate selection (Oliver, 2013; Webster and Oliver, 2007).

2.4 Digital Soil Mapping

2.4.1 Principles

DSM, as defined by McBratney *et al.*, (2003), is the creation of spatial soil maps that infer the spatial and temporal variations of soil types and soil properties from observation and related environmental variables. DSM utilises statistical models together with covariate data which could be derived from sources such as remote sensing, digital elevation models, climate models, yield monitors and geological maps to create accurate maps.. ML refers to models that use data to learn patterns in order to make predictions – a full discussion thereof is provided in Section 2.5. Internationally, this method has been successfully and commonly used in countries such as Australia (Kidd *et al.*, 2020) and the United States (Adeniyi *et al.*, 2024). It has been found that over the last two decades, the use of DSM has become more prevalent at a number of scales, including, but not limited to, field scale, regional scale and national scale (Wadoux, *et al.*, 2020). DSM maps are used extensively in PA, including, but not limited to, site-specific variable rate irrigation (O'Shaughnessy *et al.*, 2015). Mapping soil properties such as soil organic carbon content (Venter *et al.*, 2021; Cloete *et al.*, 2025) and soil organic carbon stocks (Cloete *et al.*, 2025). The industry standard method of soil mapping is largely dependent on soil sampling, classification, and property testing in laboratories. However, using ML can help increase the accuracy of predictions between sampled points. DSM utilises the SCORPAN model (McBratney *et al.*, 2003), which is based on the CLORPT soil formation model defined by Jenny (1941), which is expressed as in Equation 3, where S represents the Soil, cl the climate, o the organisms, r the relief, p the parent material and t represents time.

Equation 3: CLORPT Model

$$S = f(cl, o, r, p, t)$$

The SCORPAN model considers soil forming factors, namely, Soil (S), Climate (C), Organisms (O), Relief (R), Parent Material (P), Age (A) and Spatial Position (N). By making use of the CLORPT and SCORPAN model, it is possible to make maps that take into consideration all of the variables that can and will affect the soil. These factors can all be taken into consideration when conducting DSM by including them as covariates. Covariate data is derived from imagery (i.e., remote and/or proximally sensed) and used to make more accurate predictions compared to

predictions made without taking into account covariate data, such as OK by considering more factors regarding and influencing the specific soil property.

2.4.2 Advantages and Disadvantages of Digital Soil Mapping

DSM does have several disadvantages that might affect map accuracy. A major disadvantage of DSM is that, when conducting mapping using more complex models such as Cubist, it requires larger and more representative datasets for map creation (Arrouays et al., 2020). Although some ML models are able to make predictions on small datasets, ensemble methods like RF and Cubist are more effective with larger datasets that are generally more spatially representative. (Wadoux et al., 2020). Furthermore, DSM is completely dependent on environmental covariates, which may prove to be an issue for practitioners as some areas may lack the necessary covariate data (McBratney et al., 2003). At field scale this could be a limiting factor, as covariates which correlates with soil properties at field scale could not be available. An advantage of DSM is that previous studies have shown that DSM techniques have created more accurate maps compared to conventional mapping methods, such as OK (Hempel et al, 2008). Additionally, when utilising ML models in conjunction with DSM, it is possible to make predictions on a number of different scales, from national scale, down to field scale (McBratney et al., 2003).

2.4.3 Previous Studies in Southern Africa at Field Scale

A recent study by Cloete et al. (2025), in which ML methods were compared to OK when mapping soil organic carbon stocks (SOCS) has found that utilising ML in DSM have resulted in more accurate maps when compared to utilising the conventional method of OK at farm scale, with the DSM with ML methods, particularly RF and Cubist, achieving lower RMSE and higher MEC values. A study by du Plessis et al. (2020) on gully erosion in the Eastern Cape Province found that the use of DSM with ML was effective at producing soil maps that were used to make better informed decisions regarding the mitigation of gully erosion. This further proves the practical capability of DSM with ML in mapping soils for environmental management. Another study conducted in the Sandspruit catchment in the Western Cape Province in South Africa found that, at farm-scale, covariate selection was the most significant aspect to consider when conducting DSM, achieving this by making use of the Least Absolute Shrinkage and Selection Operator (LASSO) in order to select features when making predictions (Flynn et al., 2022).

At field scale, DSM with ML was utilised in a study in the German loess belt (Heil et al., 2022), in which an unmanned aerial vehicle (UAV) was used to gather colour imagery which was then used to make predictions of soil organic matter. The study utilised the RF ML model, which proved to be highly effective, presenting a RMSE of 0.13% for soil organic matter. The researchers concluded that the use of both the UAV imagery and the DSM model proved to be effective at mapping and monitoring soil organic matter at field scale. A study conducted by Louw (2025)

conducted using 2 fields of differing sizes, with one field being 126 ha and the other 44 ha, used Cubist, RF and OK. The different soil properties tested presented results that showed that all methods were too inaccurate for use in PA.

2.5 Machine Learning Models

ML models can be divided into three broad categories, namely, Supervised Learning, Unsupervised Learning and Semi-Supervised Learning (Sarker, 2021). Supervised learning is conducted by training a model using datasets where the property of interest is already measured, allowing the model to learn to predict the property using associated predictor variables (Nasteski, 2017). Thereafter, it can make predictions on unseen data. Examples of supervised learning methods include Linear Regression, Decision Trees (which can also be applied in unsupervised contexts), Support Vector Machines and RF models. Unsupervised learning occurs when a model is trained on an unlabelled dataset, where the model finds the hidden structure within the dataset. Examples of unsupervised learning models are k-means clustering, Principal Component Analysis (PCA) and Independent Component Analysis (ICA) (Naeem *et al.*, 2023). Semi-supervised learning combines a small, labelled dataset with a larger unlabelled dataset during model training, using the labelled data to guide the model while leveraging the structure present in the unlabelled data to improve performance (van Engelen and Hoos, 2020). In the context of DSM which mainly uses supervised learning methods, common ML models include the Cubist and RF models (Kuhn *et al.*, 2023; Breiman, 2001).

2.5.1 Cubist

The Cubist ML model is a rule-based model tree in which decision trees are combined with linear regression in order to make predictions (Kuhn *et al.*, 2023). Demir and Sahan (2024) found that the Cubist model was effective at capturing non-linear relationships due to the fact that the model constructs a decision tree based on the dataset, whereafter each leaf in the decision tree is fitted with a linear regression model, thus allowing the model to handle complex datasets.

2.5.2 Random Forest (RF)

The high variation of individual decision trees is addressed by the RF model, which is an ensemble technique. For instance, small changes in the dataset can result in vastly different decision trees. This is avoided when using RF because as it makes use of a bootstrap sample from the training data and a randomly selected subset of predictors, which are then averaged. This increases the stability of predictions and lowers variation (Breiman, 2001). It is through the construction of the large amount of decision trees that the RF model reduces variance that would normally be found when using a single decision tree by averaging across the various decision

trees. The decision trees are trained using random subsets of the data, where bootstrap sampling is used to select the data that is used to train the trees

2.6 Model Validation

2.6.1 Cross-Validation

Cross-Validation is a technique used to test the prediction performance of a statistical model, by making use of the available full dataset. The technique involves splitting the full dataset into two separate datasets, namely, the training dataset and the test/validation dataset. K-fold cross-validation separates the full dataset into k equally sized subsets, where k is the chosen number of folds. Thereafter, the model is trained k times, keeping one fold out, which is then used as the test set, using $k-1$ as the training set. Performance metrics are then averaged across all of the separate k iterations to determine a model generalisation error. When k is set as say 10, then this would be referred to as 10-fold cross-validation. Leave one out cross validation is a form of k -fold cross validation where the k is set to the number of samples, meaning that with each iteration of the model only one sample is used as validation data. This is typically used for OK accuracy assessment.

2.7 Validation Statistics

There are several validation statistics that can be utilised when determining map accuracy. These statistics include root mean square error (RMSE), coefficient of determination (R^2) and Lin's Concordance Correlation Coefficient (CCC) (Lin, 1989). These statistics could be applied with all datasets, the leave one out cross validation, 10-fold cross validation or the training and validation datasets.

2.7.1 Root Mean Square Error (RMSE)

RMSE works by measuring the average squared difference between the predicted points and the actual value of the points (Hodson, 2022). It is thus that, after calculating the RMSE, a lower value for the RMSE will be more accurate than a higher value. Equation 4 is used to calculate RMSE.

Equation 4: RMSE

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n_{val}} (obs_i - pred_i)^2}$$

where n_{val} represents the size of the dataset, obs_i represents the observed values, and $pred_i$ represents the predicted values.

2.7.2 Model Efficiency Coefficient (MEC)

The Model Efficiency Coefficient (MEC) is a statistical measure that measures the relationship between the observed and predicted values of the model. For calculating MEC, Equation 5 is used.

Equation 5: MEC

$$MEC = 1 - \frac{\sum_{i=1}^{n_{val}} (obs_i - pred_i)^2}{\sum_{i=1}^{n_{val}} (obs_i - \overline{obs})^2}$$

where n_{val} is the size of the validation dataset, obs_i is the observed value, while $pred_i$ represents the predicted value and $\overline{obs}$ is the average of the observations.

2.7.3 Lin's Concordance Correlation Coefficient (CCC)

Lin's Concordance Correlation Coefficient (CCC) is a measure of the fit between the observed and predicted values on a 1:1 line (Lin, 1989). CCC is calculated through the use of Equation 6 (as seen below). A CCC of 1 indicates perfect fit between the observed values and predicted values.

Equation 6: CCC

$$\rho_c = \frac{2\rho\sigma_{pred}\sigma_{obs}}{\sigma_{pred}^2 + \sigma_{obs}^2 + (\mu_{pred} - \mu_{obs})^2}$$

where ρ is the correlation coefficient between the observed and predicted values, μ_{obs} and μ_{pred} are the averages of the observed and predicted values, while σ_{obs}^2 and σ_{pred}^2 are the variances of the observed and predicted values respectively (Malone, Minasny and McBratney, 2016).

2.8 Summary and Research Gap

While DSM with ML has proven itself in the international context, there is little literature to prove its efficacy in the South African PA context, particularly at field scale. Furthermore, the South African industry standard method of mapping, OK conducted on samples taken on a 1-hectare grid, has not been evaluated and proven to be sufficiently accurate for PA. The financial implications of these methods have also been left unevaluated. This study addresses this gap by evaluating the industry standard as well as DSM with ML, as well as assessing the financial implications of the accuracies of the methods.

CHAPTER 3 MATERIALS AND METHODS

3.1 Study Site and Site Description

The study site is a field 354 ha field used for dryland grain production located approximately ten kilometres south of the town of Orkney, within the Free State Province (Figure 2).

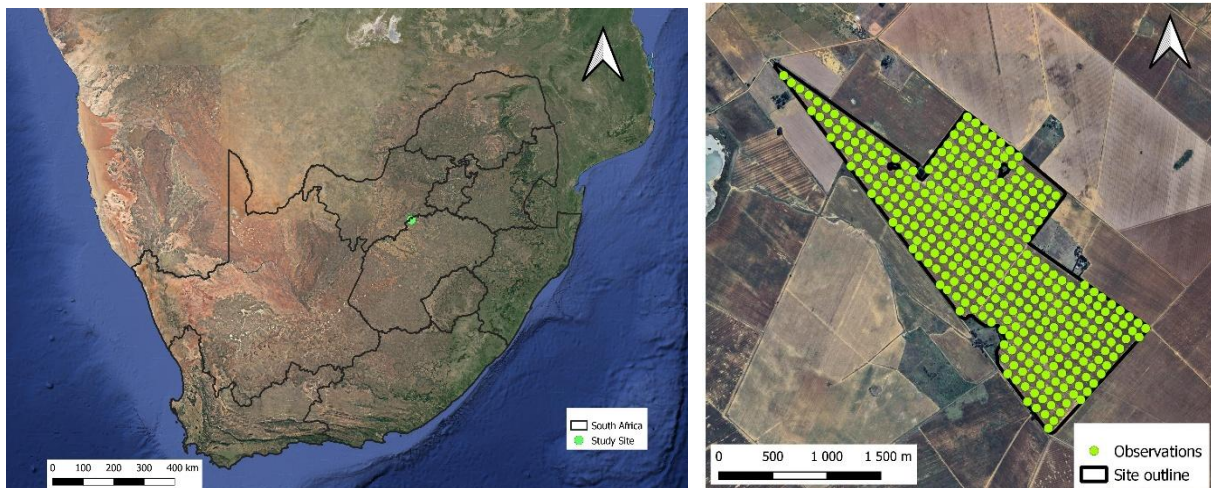


Figure 2: Location of Study Site and Sampling Points at Study Site.

The elevation varies between 1320 and 1340 metres. The geology of the area is characterised by alluvium, colluvium, eluvium and gravel (Council of Geoscience, 2019) and the study site is entirely situated upon the Bd broad land type, which is characterised as red and yellow, eutrophic, apedal soils with plinthic subsoils. The sandy, low-clay nature of the aeolian parent material has a direct influence for the relatively low CEC values found at the site. Furthermore, the parent material of the site is mainly transported material, comprised mostly of aeolian sands. The specific land type is Bd13 (Land Type Survey Staff, 1972). The area is dominated by Terrain units 1 – 3 (>70%), with the Avalon, Clovelly and Hutton soil forms predominantly associated with this land type (Land Type Survey Staff, 1972). The Median Annual Precipitation (MAP) in the area is 490 – 530 mm, while the Mean Annual Temperature (MAT) in the area is 17.3°C (Schulze, 2007). The cultivation practices in the field are mainly comprised of monoculture with maize, rotated with soybeans every fourth year. The tillage practices are largely conventional tillage. Manure is applied to the field prior to planting as well as conventional chemical fertilisers through the growing season. In 2022 soybeans were planted, while maize was planted in all the other seasons for which yield data was provided.

3.2 Soil Sampling, Data Collection and Analysis

Three hundred and fifty two (352) topsoil samples (0-20 cm) were collected using a soil auger on a 100 m grid system (1 hectare grid size) (Figure 2) and was sent for laboratory analysis for pH with the KCl method, plant available phosphorus (P) with the Bray method and exchangeable base cations, potassium (K), sodium (Na), calcium (Ca), and magnesium (Mg), with the NH_4OAc method. CEC was calculated as the sum of the base cations and the exchangeable acidity. Four consecutive years of yield data was retrieved directly from the harvester, for the years 2021, 2022, 2023 and 2024. The 2022 data for the soybean yield was not utilised in this study as it showed some clear errors, which is attributed to incorrect calibration of the harvester for soybeans.

3.3 Evaluation of Conventional Method of Soil Mapping

In order to evaluate the conventional mapping with OK, semi-variograms were constructed for each individual soil property in R Version 4.3.2, using the “fit.variogram” function within the “gstat” package (Pebesma, 2004). The industry standard method of creating maps does not include the examination of the created semi-variograms due to the utilisation of autofit functions, which produces a finished map without the post-fit inspection of the variogram. The examination of variograms after fitting allows for the avoidance of poorly fit models and improves prediction accuracy. In order to determine which semi-variogram model to utilise, a number of models were tested, namely, Spherical, Gaussian, Exponential and Stein’s Parameterization of the Matérn model. The semi-variogram model chosen for this study was Stein’s Parameterization of the Matérn model (Stein, 1999), selected due to it presenting the best accuracy statistics based on RMSE. Model selection based on lowest cross-validation RMSE was conducted independently for each of the seven soil properties. Stein’s Parameterisation of the Matérn model was found to produce the lowest RMSE for every property tested and was therefore used consistently across all seven semi-variograms.

Semi-variogram Modelling was conducted in a similar manner as Cloete et al. (2025) conducted in a study of mapping carbon at field scale. All of the sampled points ($n = 352$) were utilised, as would be the case in conventional industry procedures when mapping soil properties. Thereafter, map accuracy for each soil property were evaluated using 10-fold cross-validation. Data leakage was avoided during the 10-fold cross-validation process as each point that was withheld was predicted only on the remaining point. The statistics used to evaluate the OK maps were RMSE, MEC and CCC.

3.4 Comparison of the Conventional Soil Mapping Method and alternative mapping techniques

3.4.1 Digital Soil Mapping

Digital soil maps for the seven properties were generated using the sampled data and several covariates, such as the Saturation Index (SI), Brightness Index (BI), Redness Index (RI), Normalised Difference Vegetation Index (NDVI) (Table 2). These covariates were derived as raster files from satellite images collected from Sentinel Hub (Sentinel Hub, 2024) on the dates specified in Table 1. Each year included imagery from a wet and dry period of the year, which is included in Table 1. Furthermore, yield data was collected through the farmers harvester for the years 2021, 2023 and 2024 and were converted to raster files in SAGA (Conrad *et al.*, 2015). Examples of maps of the covariate data can be found in Figure 3, with the 2021 dry NDVI map, 2024 maize yield map, and the DEM of the site found in order from left to right.

Table 1: Satellite Imagery Capture Dates

Year	Wet/Dry Season	Date
2021	Wet	24-04-2021
2021	Dry	20-12-2021
2022	Wet	25-03-2022
2022	Dry	20-12-2022
2023	Wet	14-05-2023
2023	Dry	14-01-2023
2024	Wet	08-05-2024
2024	Dry	19-01-2024

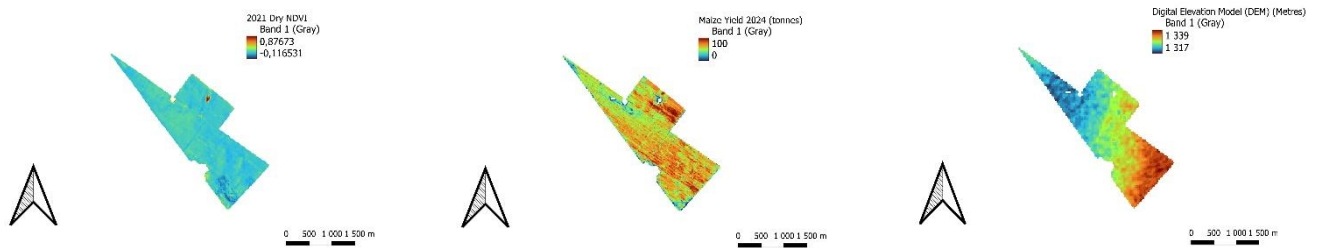


Figure 3: Examples of Covariate Data. NDVI, Yield and DEM Maps.

The Shuttle Radar Topography Mission DEM was retrieved from Earth Explorer (Earth Explorer, 2024). Thereafter, the DEM was imported into SAGA 9.5.1, where the Basic Terrain Analysis tool was used to obtain a number of other covariates, namely, the Aspect, Channel Network Base Level, Channel Network Distance, Convergence Index, LS-Factor, Plan Curvature, Profile Curvature, Relative Slope Position, Slope, Topographic Wetness Index, Total Catchment Area and the Valley Depth.

Table 1: Covariates determined from satellite images (as derived from Flynn et al., 2019)

Bands	Wavelength (μm)	Band Number	Symbol
Blue	0.45 – 0.51	2	B
Green	0.53 - 0.59	3	G
Red	0.64 – 0.67	4	R
Near infrared (NIR)	0.78 – 0.9	8	NIR
Indices	Equation	Property	
Saturation Index	$(R-B)/(R+B)$	Spectral Slope	
Redness Index	$(R^2/G*B^3)$	Hematite	
Brightness Index	$(R^2+G^2+B^2)/3^{0.5}$	Reflectance	

Normalized Difference Vegetation Index (NDVI)	(NIR-R)*(NIR+R)	Vegetation
--	-----------------	------------

The RF and Cubist models were used for this study. Both the Cubist and RF were optimised through their respective hyper-parameters, namely, committees ($C = 10$) and neighbours ($K = 2$) for the Cubist model, and ntree ($B = 50$) and mtry ($m = 2$) for the RF model. A range of values for the hyper-parameters were tested for each model in order to calibrate each respective model. The values tested for neighbours ranged from 0 – 9, while the range tested for committees was from 10 – 100 in increments of 10. A total of 23 covariates were used as predictors in the RF model. The range tested for mtry was between 2 and 10, bracketing the commonly recommended default of $\sqrt{p} \approx 4.8$, where p is the number of predictors, while ntree values tested were between 10 and 100 in increments of 10. This range was sufficient to demonstrate that the out-of-bag (OOB) error stabilised well before the upper end of the tested range, indicating that additional trees (including the software's default of 500) would not have materially improved model performance, while incurring unnecessary computational cost. The optimal values were selected based on lowest RMSE. MEC and CCC were calculated as additional validation statistics but were not used in the selection process.

Both the Cubist and RF models were trained and used to generate spatial predictions in R using the “caret” package (Kuhn, 2008) with the final predictions being made across the full spatial grid using the selected hyperparameter configurations. The full sampled dataset on the 1-hectare grid (Figure 2) was split into a training and validation dataset, consisting of 70% ($n_{\text{train}} = 246$) training data and 30% ($n_{\text{test}} = 106$) validation data. Hyperparameter tuning was conducted exclusively on the training dataset. The split was performed once at the outset of the analysis, with the training and validation datasets saved as separate files prior to any model training or hyperparameter tuning. Hyperparameter tuning was conducted exclusively on the training dataset via internal resampling within the “caret” package (Kuhn, 2008), the validation dataset. It is acknowledged that a single random split such as was used in this study can produce a performance estimate that is sensitive to the particular split obtained in the process. Repeated splits or k-fold cross-validation would provide a more robust estimate of model performance and are recommended for future work. The mapping methods were evaluated using validation methods that differed due to methodological necessity, as removing points from the dataset before conducting OK would distort OKs already data-limited spatial structure estimate. It is thus that each of the mapping methods was evaluated independently and not compared against one another, which is an

acknowledged limitation of the study. OK was assessed using 10-fold cross-validation using all of the 352 sampling points, while the ML methods were evaluated based on a simple split dataset. Therefore, the methods were not directly compared, rather, each method is assessed based on its own performance. Furthermore, graphs of the observed vs predicted values of the ML maps were generated using strictly out-of-sample predictions, made by applying the trained models to the validation dataset, which was then plotted against the corresponding observed validation values. These graphs were created using the “ggplot” function from the “ggplot2” package in R (R Core Team, 2020).

3.5 Quantification of Financial Implications of the Mapping Methods for Fertiliser Recommendations

3.5.1 Fertiliser Application Maps

The financial implications of fertilisation recommendations based on the conventional and DSM methods were calculated for phosphate (P) and potassium(K). Fertiliser recommendation maps were made for the field using the OK and DSM with ML predicted maps. Both sets of maps were added to a QGIS project (QGIS Development Team, 2023) where they were vectorised using the “raster to vector” tool. The “Dissolve” tool was used to dissolve the maps into groups with similar property values, based on the respective increment values for the given soil properties as given in the Fertiliser Handbook (FERTASA, 2016). These included 0.5 units for pH, 20 mg.kg⁻¹ for K and 5 mg.kg⁻¹ for P and 1 tonne.ha⁻¹ for yield, respectively. Fertiliser recommendations for P and K were made based on a yield potential map, the P map, the K map and the pH map respectively. The P, K and pH maps were imported into ARC GIS (ESRI, 2020), where the Union tool was used to merge each of the ML and Kriging P, K, and pH maps with the Yield Potential map in order to sort the field into zones according to the P, K or pH maps in relation to the Yield Potential in the field. The aforementioned maps were then imported back into QGIS (QGIS Development Team, 2023) where the areas were assigned a fertiliser recommendation value according to the Fertiliser Handbook (FERTASA, 2016).

3.5.2 Financial Simulations

Financial simulations were conducted by setting a price for maize, phosphorus and potassium products based on market prices as at 01/03/2024. The maize price was set at R6000 based on the SAFEX WMAZ market price on 10/02/2025. For phosphorus a commercial product with 18% P was selected, while KCl containing 50% K was selected as K source. The prices are given in Table 3.

The soil class map with its accompanying yield potential values which was created by an agronomist was supplied by the farmer. The average yield estimate was used as the expected

yield for that mapping unit. The average yield was based on a multi-year average of the yield, which does not account for season-specific factors, such as rainfall variations. The under-fertilisation costs calculated in this study must therefore assume that any shortfalls to the average are attributable to the fertiliser status. Fertiliser recommendations for each mapping unit were made based on the mapped P and K values, with the Kriging and DSM methods respectively, while aiming for a yield target of the expected yield for that mapping unit. Each of the 352 points, which represent 1 ha each, was treated as an independent evaluation unit. Therefore, each point was given a yield potential value based on the yield potential map zone that it fell in. Thereafter, the soil property measured value (P, K) at that point was used to make a determination of the required amount of fertiliser at that point based on the yield potential.

Over and under fertilisation was calculated for each sampling point, and each point was assumed to represent 1 ha. The amount of fertiliser recommended was subtracted from the amount of fertiliser which should have been applied based on the respective laboratory analysis values. Overfertilisation was noted when the fertiliser recommended based on the map was more than what should have been recommended based on the analysis results. In this case the extra costs for the fertiliser were calculated as seen in Equation 7:

Equation 7: Extra Fertiliser Cost

$$\text{Extra Fertiliser Cost} = (\text{Prediction Fert} - \text{Optimal Fert}) * \text{Fertiliser Price} * \text{Area}$$

The cost of under fertilisation was assumed to be the lack of yield achieved. The yield based on the Kriging and DSM predictions was determined by creating a graph of the two values between which the recommended value falls into on the x-axis, with the two yield target values on the y-axis, as shown in the example in Figure 4. This was then used to determine a regression equation, which was then used to determine the yield that would have been produced at the amount of fertiliser recommended by the Kriging and DSM maps. The equations differed greatly depending on the different combinations of values and thus are not included in the thesis due to the large number of equations.

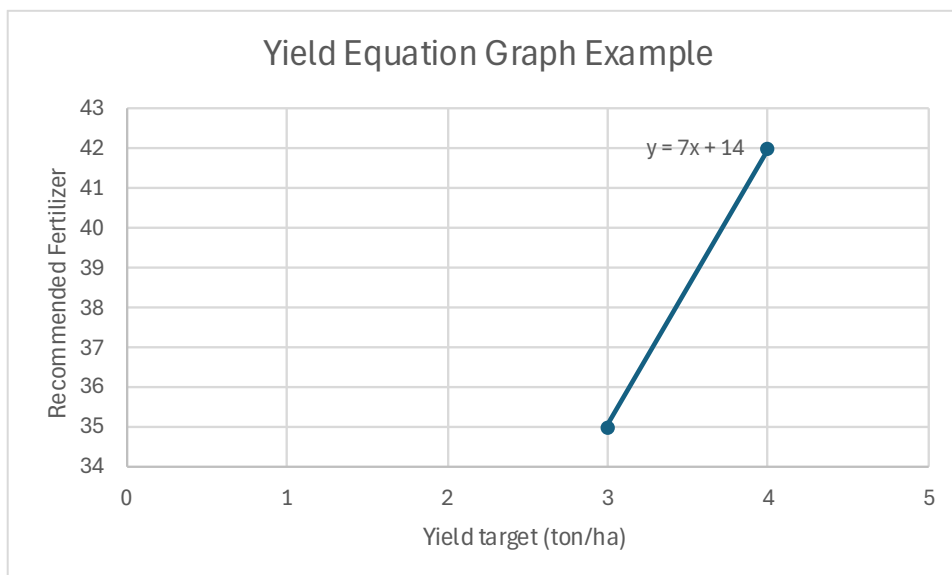


Figure 4: Example of a graph used to determine the yield equations.

For each scenario, the total income from the field was calculated and the fertiliser cost subtracted, which gave the financial implications of the different mapping units. It should be noted that additional costs, such as land cultivation, herbicide, pesticide, fungicide and fertilisation other than P, K and liming was not considered, as these are not dependent on the mapping method. The profit made from this field will therefore be much lower than what is reported on here.

The total P and K requirements for the field were calculated by multiplying the area of each zone with the Phosphorus and Potassium application recommendation. Thereafter, the total requirement per each of the soil properties was multiplied by the price per ton of each fertiliser (Table 3). This process was repeated for each of the soil mapping methods in order to determine the fertiliser cost of each of the methods.

Table 2: Fertiliser Price Table

Fertiliser Product	Price (Rand/ton)
KCl (50% K)	16470.00
Phosphate source (18% P)	10460.00

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Introduction

This section presents the main results of the study, including (i) the maps created through the two aforementioned mapping methods, (ii) the comparison of the mapping methods both visually and statistically, as well as (iii) the fertiliser recommendations based on the maps and the cost analysis of each of the methods.

4.2 Descriptive Statistics

The soil properties of the study site are generally within range for arable land in South Africa (FERTASA, 2016; Table 4). The pH of the soil falls between the minimum of 3.4 and maximum of 8.4, indicating that certain parts of the field are very strongly acidic, and others being moderately alkaline (Fichtner et al., 2019). This particularly wide range in pH values can be attributed to historical liming practices across the field, as historical fence lines would not have been treated with lime. The mean pH of the soil is 5.01, which falls within the optimal pH level for cultivation of *Zea Mays L.* which requires pH levels of over 4.5 (Department: Agriculture, Land Reform and Rural Development, 2022). The standard deviation of the pH is 0.83.

The median phosphorus (P) value in the field is higher than normal at 28.00 mg/kg and is described as “Above Optimal” in the Fertiliser Handbook (FERTASA, 2016). This shows that the field is well fertilised, through the manure application which the farmer routinely does on the field. Manure application is known to increase the P content in soils (Qin *et al.*, 2020). The minimum and maximum P values are 9 mg/kg and 82 mg/kg respectively. The standard deviation of the P values is 11.55 mg/kg.

With regard to the potassium (K) values in the field, the values vary considerably. The median value of 50.00 mg/kg is well above the lower threshold level of 29 mg/kg as defined in the Fertiliser Handbook (FERTASA, 2016). The minimum and maximum values of K in the field are 12 mg/kg and 322 mg/kg respectively. The standard deviation of the K values in the soil was 40.16 mg/kg. Despite the large variation, there are only a small number of values above the lower K application threshold of 80 mg/kg, which shows that the field will require K fertilisation.

The median magnesium (Mg) value in the field is 44.00 mg/kg. The range of the Mg values in the soil is high, with a minimum Mg value of 7 mg/kg and 284 mg/kg. The standard deviation of the Mg in the soil was 27.71 mg/kg. The distribution of the points shows a slight positive skew, with a number of higher values on the upper end of the range. The median value of the field is slightly below the normal value for sandy soils in South Africa, which is 50 mg/kg (FERTASA, 2016). Therefore, this indicates that at least half of the values are below the norm value for South African soils.

Table 3: Descriptive Statistics of Soil Properties.

Property	Mean	Median	Min	Max	Range	SD
pH	5.01	4.90	3.40	8.40	5.00	0.83
P (mg/kg)	29.56	28.00	9.00	982.00	73.00	11.55
K (mg/kg)	59.74	50.00	12.00	32.00	310.00	40.16
Mg (mg/kg)	48.70	44.00	7.00	284.00	277.00	27.71
Na (mg/kg)	6.99	3.00	1.00	259.00	258.00	19.94
CEC						
(cmol(+))/kg)	1.57	1.30	0.60	6.60	6.00	0.80
Ca (mg/kg)	175.24	147.00	21.00	921.00	900.00	109.50

Sodium (Na) levels in the field range between a minimum of 1 mg/kg and a maximum of 259 mg/kg. The higher Na values in the field are due to a pond in the field. The mean Na value in the field is 6.99 mg/kg, and the standard deviation was 19.94 mg/kg. The median value of the sodium levels is 3 mg/kg. The distribution of the Na values is positively skewed, with a number of high values on the upper end of the range.

The mean CEC is 1.57 cmol(+)/kg, while the values range between a minimum of 0.6 cmol(+)/kg and a maximum of 6.6 cmol(+)/kg. The standard deviation was 0.8 cmol(+)/kg. The median value for CEC is 1.3 cmol(+)/kg. The CEC values have a positive skewness. Calcium (Ca) levels in the soil range from a minimum value of 21 mg/kg and a maximum value of 921 mg/kg. The mean Ca value is 175.24 mg/kg, and the standard deviation is 109.5 mg/kg. The median Ca value is 147 mg/kg. Thus, the majority of the values are higher than the lower threshold level defined in the Fertiliser Handbook (FERTASA, 2016) which is > 100 mg/kg. The data also has a positive skewness. The raw data points can be found below in Figure 5.

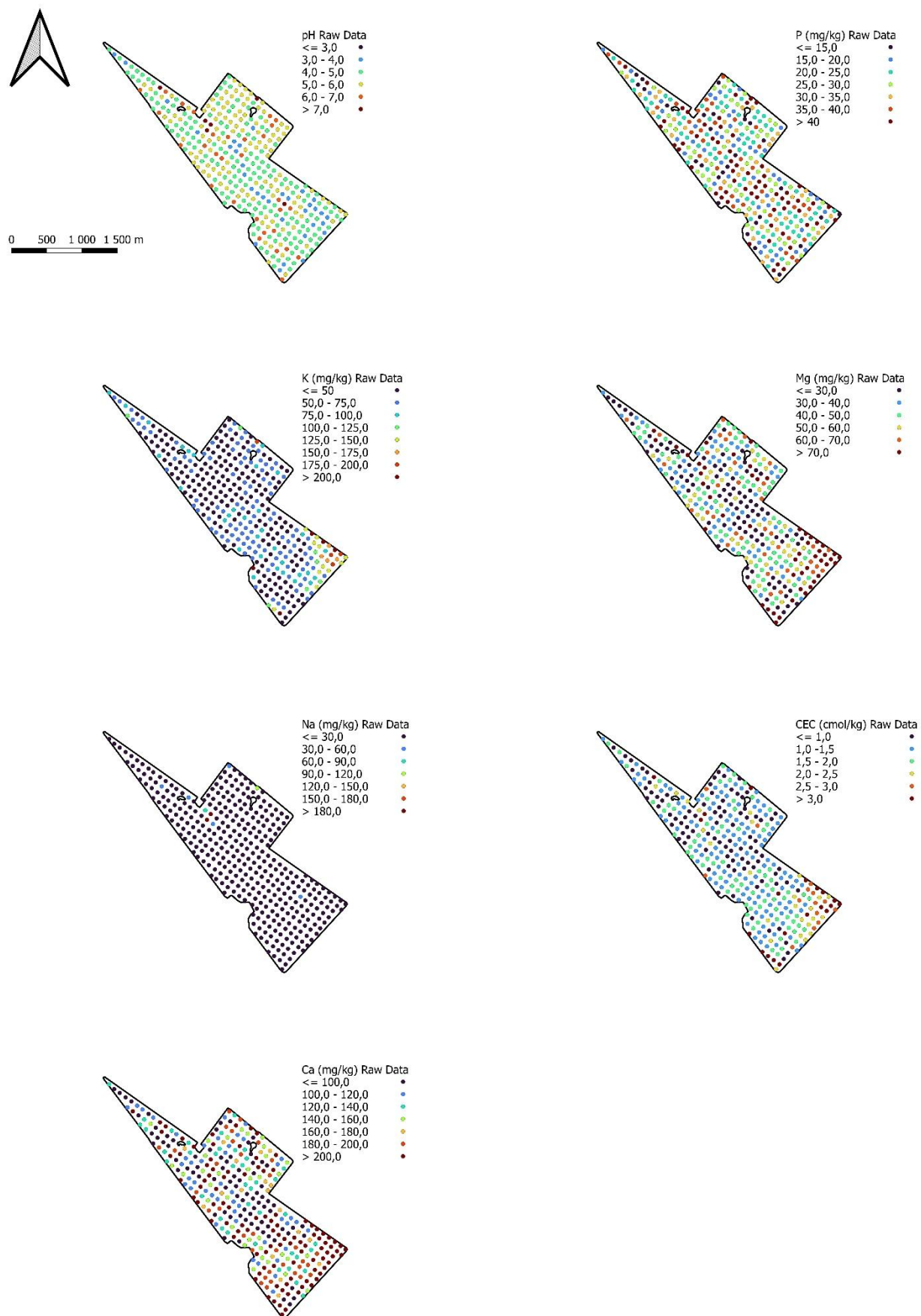


Figure 5: Raw Data Maps

4.3 Comparison of Different Mapping Methods

4.3.1 Soil pH

The pH semi-variogram and OK map show the range found for pH was approximately 250 m (Figure 6 (a)). Thus, the 1-hectare grid size used here and proposed by Verster et al. (2022) can be used in order to map the pH as the range is higher than the 100 m sampling distance employed in the 1-hectare grid system. The pH observations exhibit spatial dependence up to an estimated range of approximately 250 m.

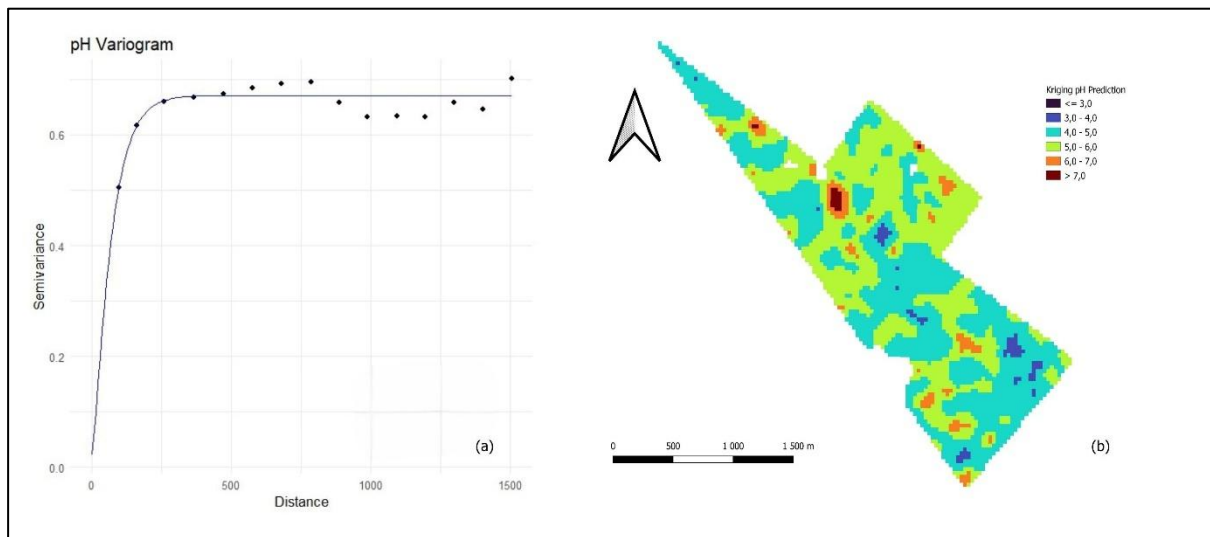


Figure 6: The constructed semi-variogram (a) and OK map (b) of pH.

The pH DSM with ML prediction maps can be found in Figure 7, as well as the accuracy statistics of the pH predictions in Table 5. With regards to the pH maps, there is a significant difference in prediction across the three predictions. It is apparent that both the Cubist and RF maps included a number of lines across the field with low pH values that were historically fence lines and breaks in between fields. It is possible that these lines were previously not limed and therefore is correctly indicated as low pH zones. Unfortunately, this cannot be verified with the current data, but it does show a possible advantage of using DSM above kriging, as such potential management outliers could be mapped. The pond found in the north of the field show higher pH values in all three predictions. When looking at the prediction statistics, the two DSM-ML methods had very similar statistics with RMSE values of 0.61 and 0.6 for Cubist and RF respectively, CCC values of 0.32 and 0.33 and MEC values of 0.13 and 0.14, while the OK had a slightly higher RMSE value of 0.77, a slightly lower CCC value of 0.19 and a MEC of 0.16. One needs to note that the Kriging map was validated using 10-fold cross-validation, meaning all datapoints were included in the mapping effort, while the DSM maps used an independent soil dataset. Thus, the validation statistics are not directly comparable due to the differing methods of cross-validation. The fact that the MEC results for all three of the predictions were so low can be attributed to the low variability of pH values in the soil. However, a low MEC indicates that the model explains little of

the observed variation relative to the mean benchmark, and this should be treated as a genuine indicator of poor explanatory performance in its own right, regardless of the RMSE result. The validation statistics for all three maps are disappointing and shows higher uncertainties than necessary for accurate soil fertility management, especially since the RMSE is above 0.5 pH units, which is typically regarded as a management unit for pH (FERTASA, 2016), thus rendering the predictions not useful for mapping pH levels. Robinson and Metternicht (2006) achieved a better result when using OK to map topsoil pH, with an RMSE of 0.2475, which thus can be considered sufficiently accurate. It should be noted that cross-study RMSE comparisons such as these are of limited value, as RMSE is influenced by factors such as the variability of the response variable, differing sampling designs and spatial resolution differences between studies. Such comparisons are therefore used as indications alone. The RMSE values for the Cubist and RF maps were lower than those presented by Louw (2025) in Field 1 of their study, while the RMSE for their OK result was lower than those in this study. Furthermore, the CCC presented far better results than Louw (2025), but were still too low to be considered accurate. The MEC results presented in this study were comparable for both the Cubist and RF results but were still too low to be considered meaningful. The OK MEC result outperformed that of Louw (2025) but is still considered to be inaccurate. However, Field 2 presented a different set of results, where the results they presented were far better for all methods in the RMSE, CCC and MEC. The likely reason that Field 2 yielded better results is due to a stronger relationship between measured soil properties and available covariates.

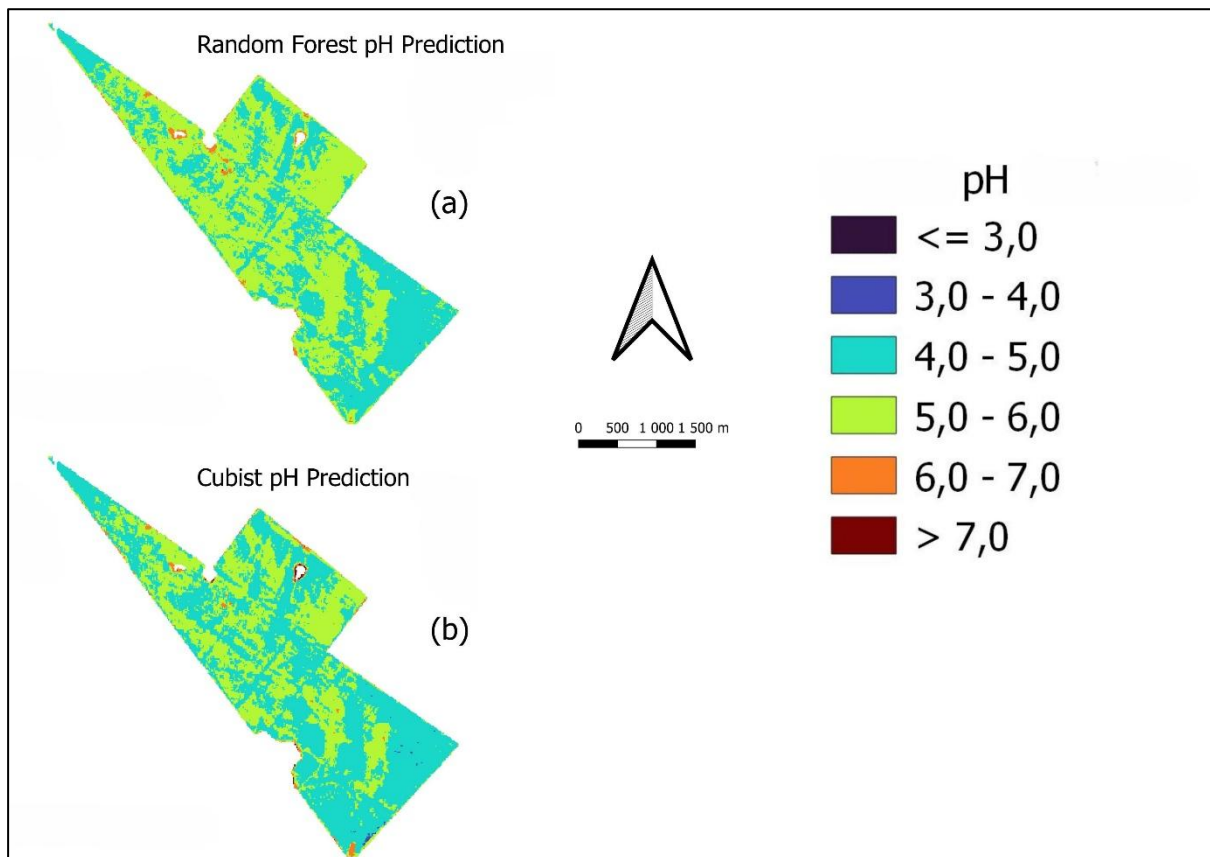


Figure 7: The generated pH maps of RF (a) and Cubist (b)..

Table 4: The accuracy statistics of the pH maps

pH Mapping Method	RMSE	CCC	MEC
OK	0.77	0.19	0.16
Cubist	0.61	0.32	0.13
RF	0.60	0.33	0.14

4.3.2 Phosphorous

Even though the most optimal hyper-parameters combinations were applied, the P semi-variogram did not follow the datapoints, which is indicative of a very low spatial correlation between the points (Figure 8 (a)), indicating that there is weak spatial correlation between the data points. The overly smoothed kriging map (Figure 8(b)) confirms this. The model parameters do, though, indicate that the 100-ha grid is sufficient for P mapping as the range is 349 m, as the points were sampled well within the range, at 100 m. This is however misleading, and a smaller grid is required to map P by kriging on this field. This can be seen in that the semi-variogram model does not follow the dots of the variance as can be seen in the raw data in Figure 5. The

over-smoothing of the P prediction could be due to the lack of spatial variability of the P values in the field (Yamamoto, 2005). This phenomenon highlights the need for better scrutinizing of the raw data and not just assuming the methodology followed as in the industry standard will result in accurate soil maps.

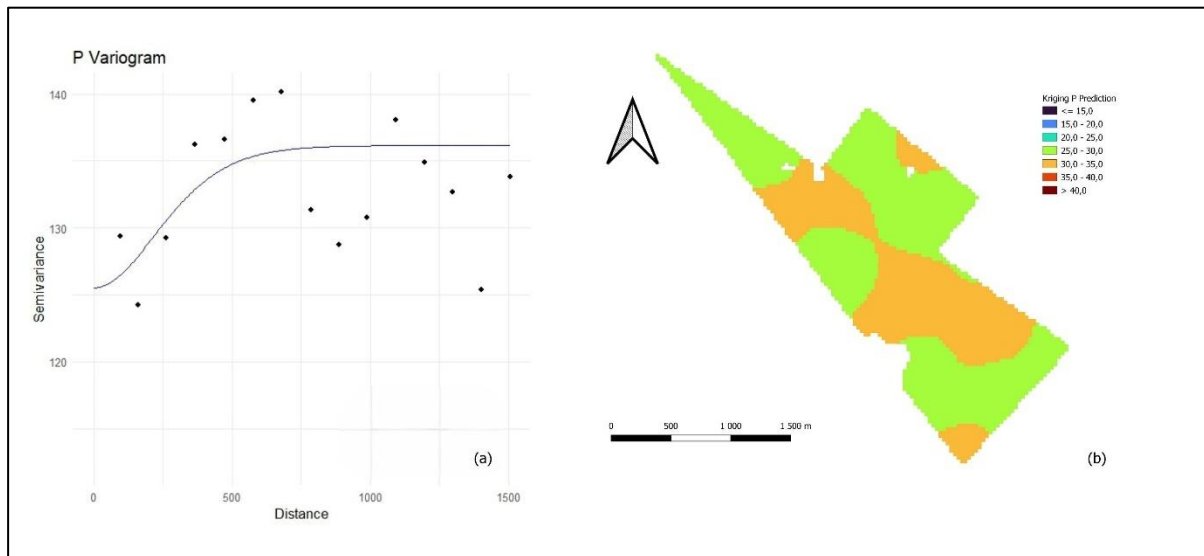


Figure 8: The constructed semi-variogram (a) and OK map (b) of the phosphorus (P).

The P prediction maps in Figure 9 show a large difference between the three methods. As mentioned, the Kriging map shows over smoothing of the data, while the Cubist map shows only a single value for the entire field (28 mg/kg). This presumably happens when an average value will return better prediction statistics than using the model. The RF map does seem to follow some form of yield map, and looks as if it can be legitimate, but unfortunately the validation statistics (Table 6) do not back this up. Despite the large difference in map appearance, the RMSE values are very similar. The RMSE values also indicate that the predictions are unusable as they are greater than the P class widths found in the Fertiliser Handbook (4 mg/kg) (FERTASA, 2016), with all of the RMSE values being close to triple the class widths. This highlights the value of not only relying on statistics for validation purposes. The CCC and MEC values, which could only be measured for Kriging and RF, does indicate that the maps are very poor and does not reflect reality at all. All of the methods proved to be ineffective at mapping the P. Overall, the industry standard method using OK showed very poor statistics and very little correlation between points. This result concurs with the findings of Louw (2025) and further proves the need for an improved method of soil property mapping in PA. The RMSE of the Cubist was better than the study by Kaya et al. (2022), who achieved a Cubist prediction with an RMSE of 15.94 mg/kg. However, there is no CCC value to make a comparison to the Kaya et al. (2022) study. Overall, the DSM with ML methods both yielded results that were poor. As was in the case with the OK, Louw (2025) had similar results in that the maps were too inaccurate for use in PA and that had low

autocorrelation between points. Interestingly, Louw (2025) also could not determine an MEC value for Cubist, as the model failed to find rules for the prediction and averaged the field as a whole, which is similar to the result presented below (Figure 9). All models failed to produce reliable prediction maps for P, which is indicated by the near-zero MEC values and maps with very little spatial coherence. There is likely a weak relationship between the P values and the covariate data. The maps are therefore unintelligible for variable-rate fertiliser applications.

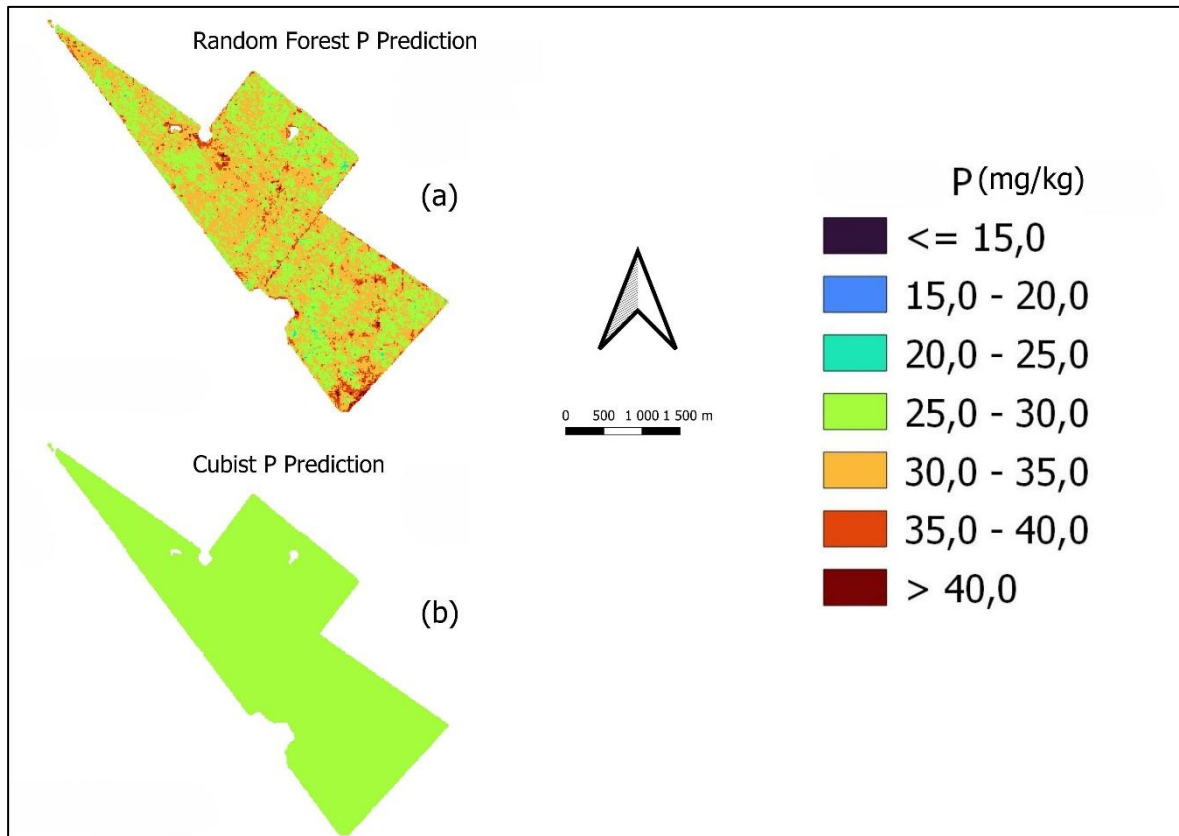


Figure 9: The generated phosphorus (P) maps of RF (a) and Cubist (b).

Table 5: The accuracy statistics of the phosphorus (P) maps.

P Method	Mapping Method	RMSE (mg/kg)	CCC	MEC
OK		11.59	0.03	0.0009
Cubist		11.13	NA	-0.002
RF		11.41	0.02	-0.04

4.3.3 Potassium (K)

The K semi-variogram (Figure 10 (a)) did not reach the sill within the maximum lag distance applied. This therefore indicates that spatial correlation continues past the extent of the study area. Thus, the sampling distance of 100 m should be sufficient. However the steady rise of the

variogram is most probably due to directionality within the data, which is also indicated in the OK map (Figure 10 (b)), with a clear gradient towards the eastern corner with much higher values. Directionality is present and can be seen when comparing the below result to the raw data in Figure 5. The directionality should be dealt with by using a different method, such as regression kriging (Zhu and Lin, 2010), but this is typically not done in industry. All three maps (Figure 11) show this trend of relatively low values and then a steep increase towards the eastern corner. The RMSE values are similar for all three methods, and are very high, especially when compared to fertiliser recommendation classes, which typically has ranges of 10 mg/kg (FERTASA, 2016). The CCC and MEC values concur. For CCC OK outperforms Cubist by a long way, which achieved better values than RF. The MEC value is also highest for OK, and then similar for Cubist and RF, albeit very low. When taking all three validation values into account, all three maps are not acceptable for fertiliser recommendations; for OK it is the high RMSE, while for RF it is the low CCC and MEC values which disqualifies the map from being acceptably accurate. Cubist shows poor accuracy with all three measures. The OK map (Figure 10 (b)) displays some over-smoothing in the northern parts of the field as a result of a lack of variability in the K values in that area, causing a weak spatial autocorrelation (Yamamoto, 2005).

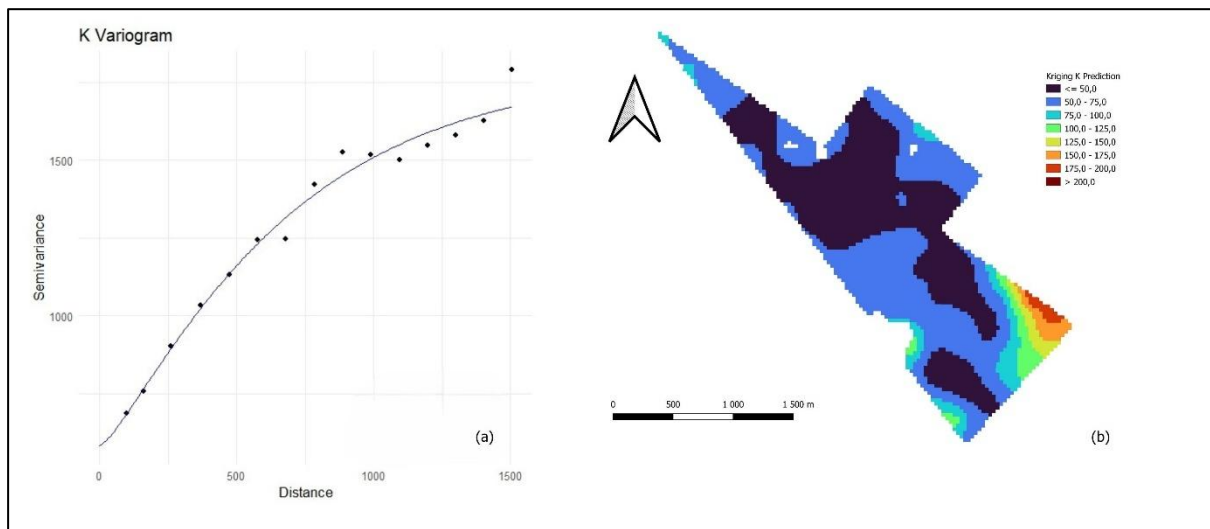


Figure 10: The constructed semi-variogram (a) and OK map (b) of the potassium (K).

The DSM-ML (Figure 11) maps show a gradual change throughout the field, with a good deal of variation through the field without any clear patchy areas. The gradual pattern observed in both of the DSM-ML maps mirrors the spatial distribution of the yield and spectral covariates that were used to train the models. The influence of the covariate data on the model prediction cannot therefore be understated. This is consistent with the covariate-dependent nature of the DSM methods (Flynn et al., 2022), where a predictions spatial structure is heavily influenced by

covariates, especially in cases where the relationship between the covariates and soil properties is weak.

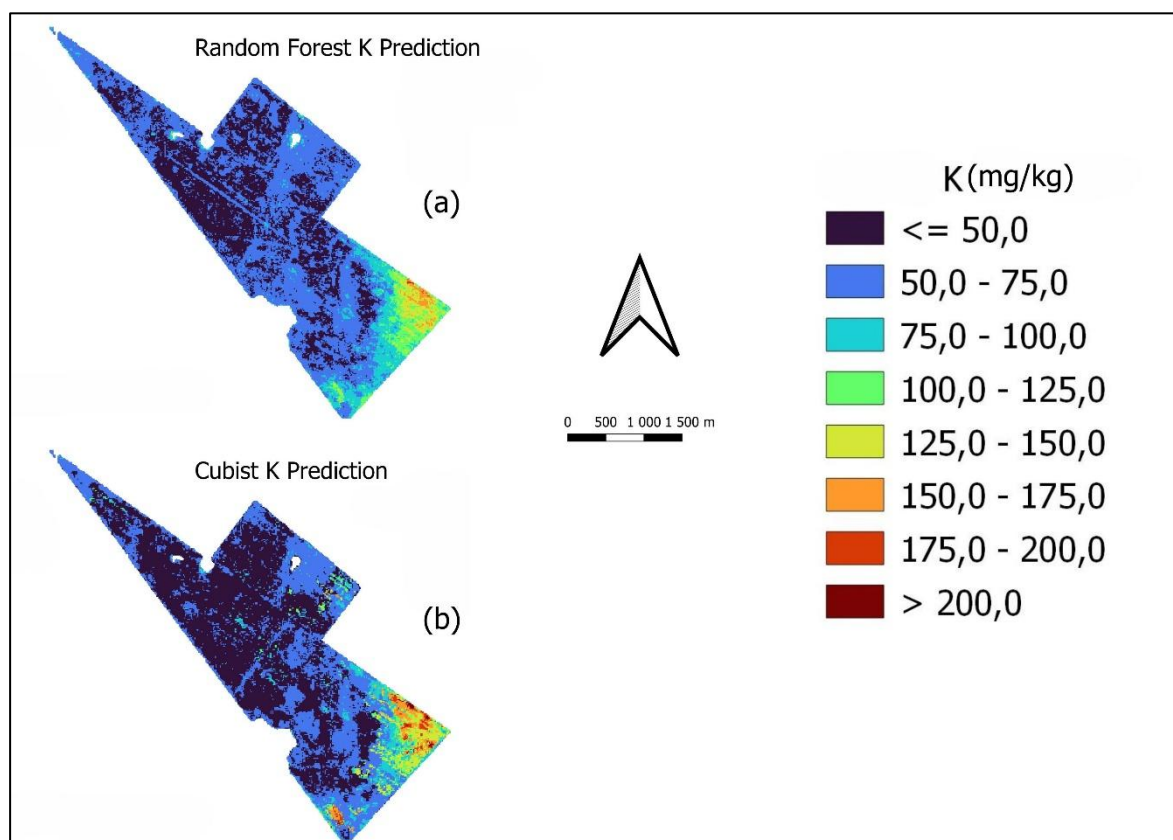


Figure 11: The generated potassium (K) maps of RF (a) and Cubist (b).

Table 6: The accuracy statistics of the potassium (K) maps.

K Mapping Method	RMSE (mg/kg)	CCC	MEC
OK	30.73	0.52	0.47
Cubist	35.22	0.32	0.18
RF	34.80	0.18	0.19

The RMSE of the methods (Table 7) fell well outside of the class widths found in the Fertiliser Handbook (5 mg/kg) (FERTASA, 2016), with the RMSE of OK being over 6 times larger than the class widths, and the Cubist and RF presenting RMSE values that were around 7 times larger than the class width. When comparing to the RMSE results of Louw (2025), the statistics of this study were far outperformed. The error presented in this study could result in the areas within the field being classed incorrectly, while those of Louw (2025) could still fall within the class ranges of 25 mg/kg as stipulated in the Fertiliser Handbook (FERTASA, 2016). The CCC results of this

study were better than those presented by Louw (2025), with the OK result performing particularly well in comparison, with 0.52 from this study and 0.25 from the results of Louw. Furthermore, the MEC of this study significantly outperformed those of Louw (2025) in every instance.

4.3.4 Calcium (Ca)

The Ca semi-variogram (Figure 12 (a)) again shows directionality, which is also indicated on the OK map (Figure 12 (b)). However, the semi-variogram again displays a large range of 3002 m, indicating that a 100-ha grid would have been adequate. Again, the need for usage of additional methods is highlighted.

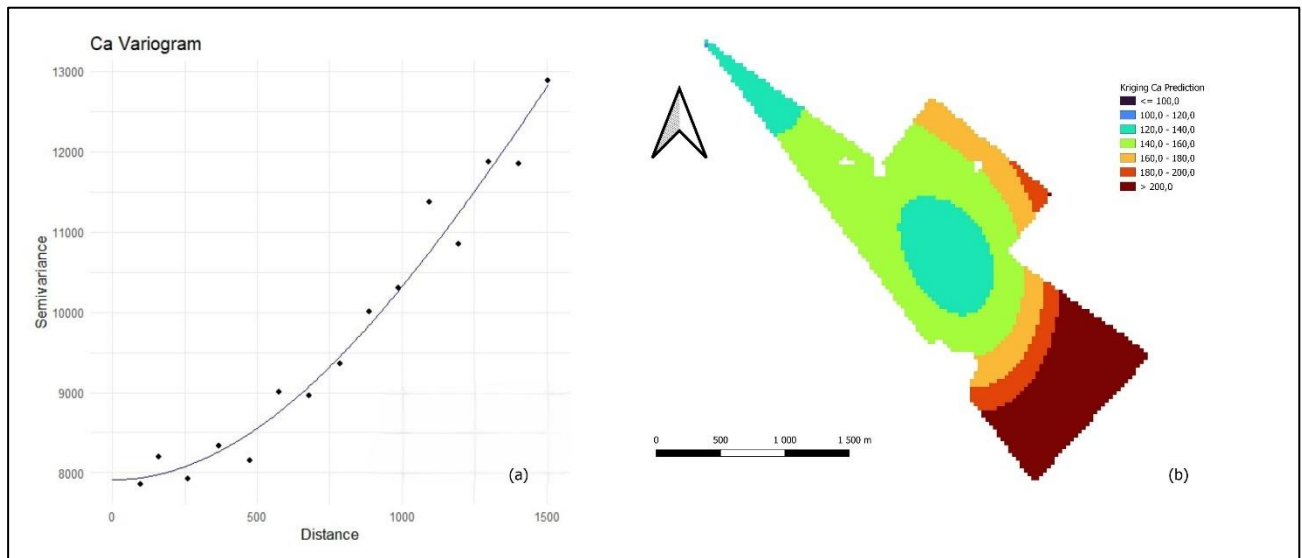


Figure 12: The constructed semi-variogram (a) and OK map (b) of the calcium (Ca).

The two Ca prediction maps found in Figure 13 have a similar distribution throughout the field, with an increase towards the southeast, already mentioned to introduce directionality. The DSM maps also indicate the old fence lines as previously seen with the pH maps. This could strengthen the case for old fence lines or roads not being limed, which would have increased the Ca content of the soil as well. When examining the statistics of the maps (Table 8), the RF and Cubist showed similar RMSE and MEC values. Despite having the highest RMSE, OK's MEC was comparable to RF's, likely reflecting Ca's highly skewed distribution, where MEC is sensitive to how well extreme values are captured, independent of average error magnitude. Cubist's strong CCC alongside its very low MEC likely reflects the same skew: CCC can remain reasonably strong based on agreement across the bulk of values, while MEC is heavily penalised by any failure to capture the extreme high Ca values that dominate the total variance. Both the RMSE and MEC values were somewhat lower than the OK values, but of course lower means better for RMSE and poorer for MEC. For CCC the accuracy was best for Cubist, then RF and lastly OK. Due to Cubist having a better CCC, and a marginally higher RMSE compared to RF, the Cubist result

was the best of the two ML results. However, based on the overall validation statistics, the values do not indicate maps which should be used for fertiliser recommendations (Louw, 2025).

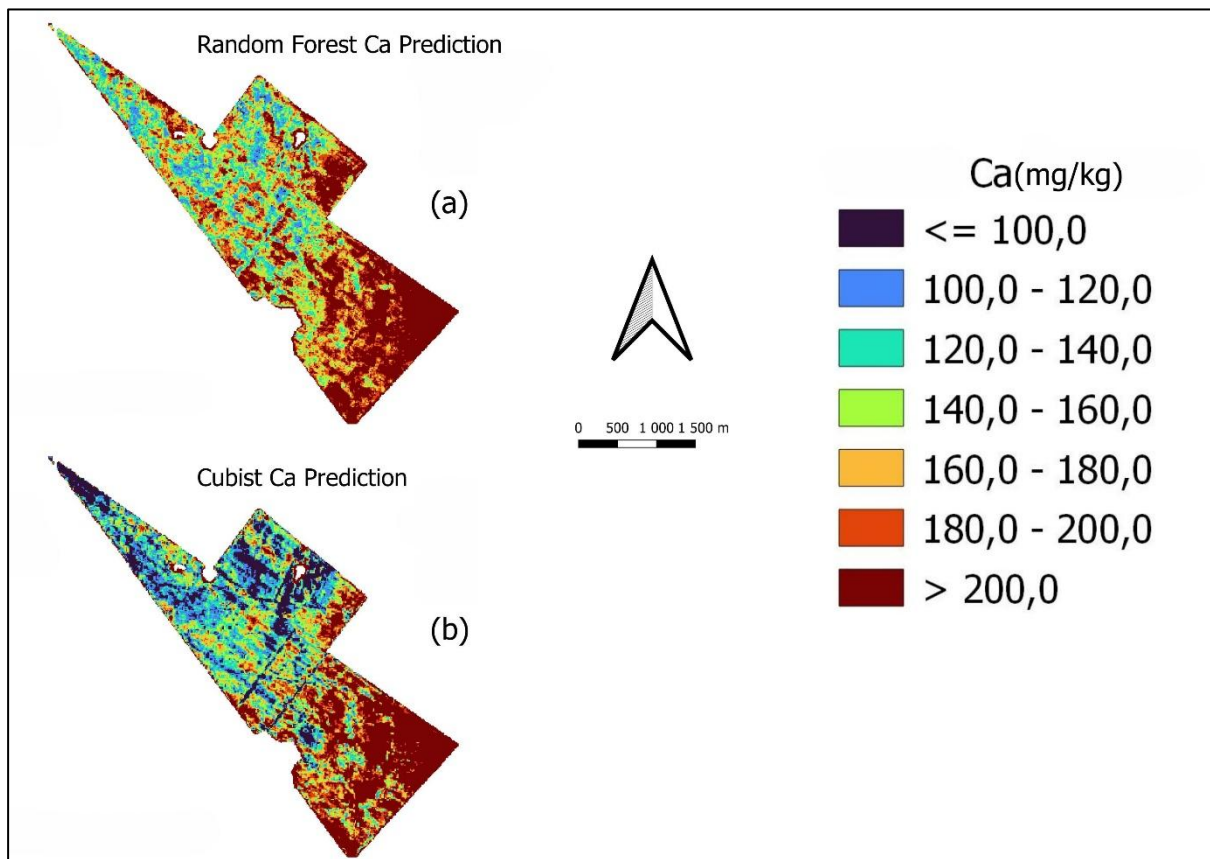


Figure 13: The generated calcium (Ca) maps of RF (a) and Cubist (b).

Table 7: The accuracy statistics of the calcium (Ca) maps.

Ca Mapping Method	RMSE (mg/kg)	CCC	MEC
OK	95.82	0.34	0.21
Cubist	70.84	0.42	0.08
RF	68.97	0.38	0.26

4.3.5 Magnesium (Mg)

The semi-variogram of the Mg (Figure 14 (a)) indicates a range of 2233.573 m, with a nugget:sill ratio of 0.77. This is indicative of a weak spatial autocorrelation, which resulted in over-smoothing of the OK Mg result (Figure 14 (b)); Yamamoto, 2005). The semi-variogram shows a similar trend to that of the K and Ca, in which there is significant directionality, therefore resulting in high values in one end of the field, and far lower values on the other end. This brings into question the accuracy and usefulness of the Mg prediction through the use of OK, as the map does not accurately represent the values in the field.

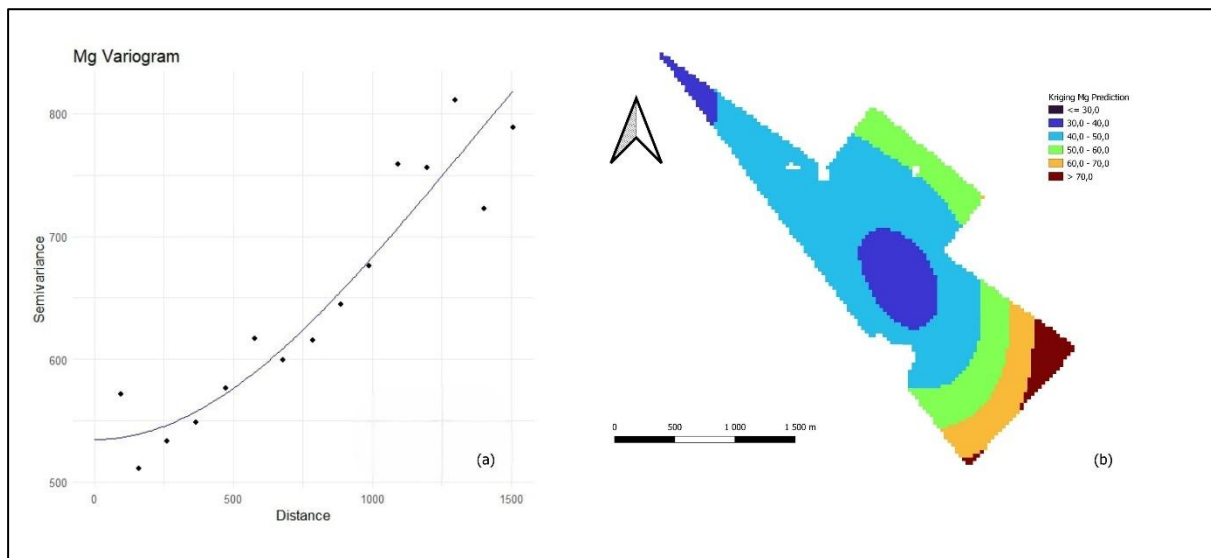


Figure 14: The constructed semi-variogram (a) and OK map (b) of the magnesium (Mg).

The OK prediction (Figure 14 (b)) performed poorly in the case of Mg, whereas the Cubist result (Figure 15 (b)) was more accurate, indicated by both the RMSE and CCC of the properties (Table 8). There are not many visual similarities between the OK prediction (Figure 14 (b)) and the Cubist prediction (Figure 15 (b)), with more similarities occurring between the Cubist prediction (Figure 15 (b)) and the RF prediction (Figure 15 (a)). Overall, the OK results were unsatisfactory, as the RMSE result would result in the error being larger than two classes in the Fertiliser Handbook (FERTASA, 2016), while the RMSE of both the RF and Cubist results were unsatisfactory as well, they were more accurate than the OK. However, the RMSE could still result in the prediction falling into another class in the Fertiliser Handbook (FERTASA, 2016). The most accurate of the three results was the Cubist result which outperformed the other two in both RMSE and CCC, but not in MEC. While the Cubist performed the best of the group, the models did not produce sufficiently accurate results for precision mapping.

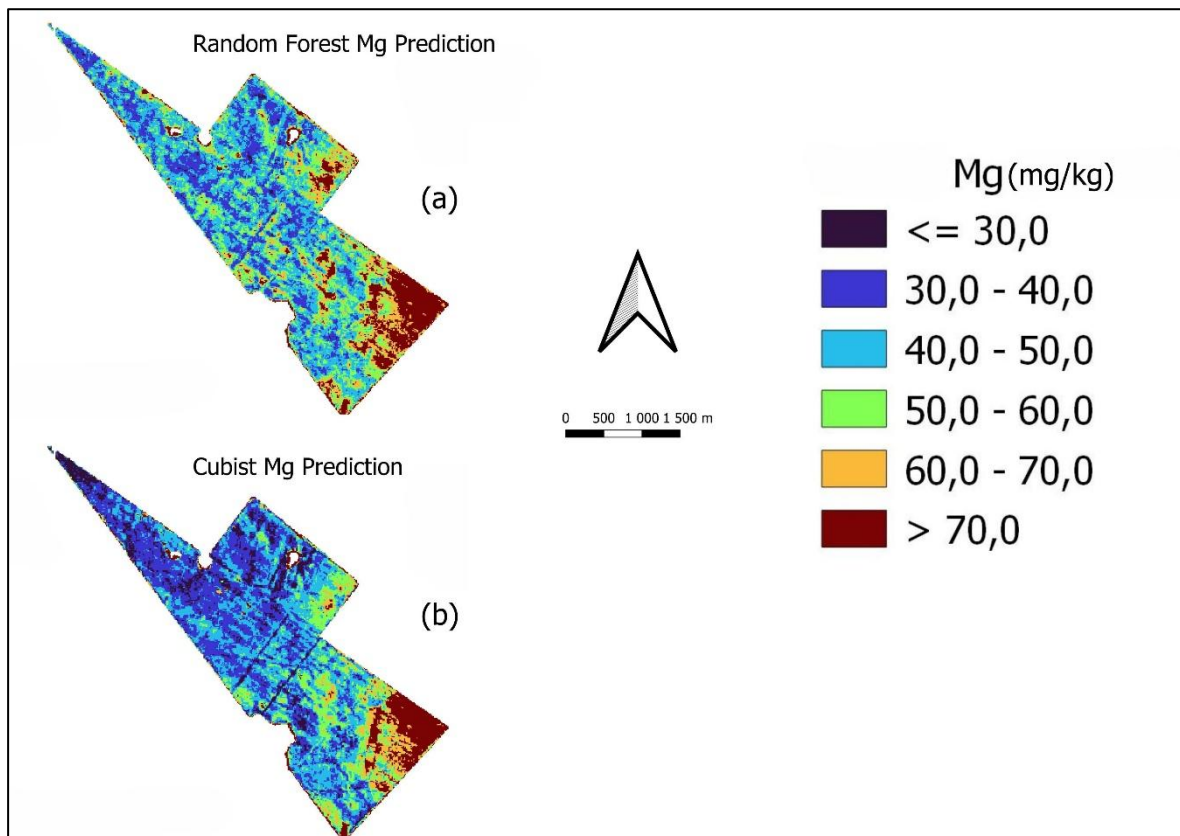


Figure 15: The generated magnesium (Mg) maps of RF(a) and Cubist (b).

Table 8: The accuracy statistics of the magnesium (Mg) maps

Mg Mapping Method	RMSE (mg/kg)	CCC	MEC
OK	25.78	0.23	0.09
Cubist	15.13	0.46	0.23
RF	16.15	0.41	0.14

4.3.6 Sodium (Na)

The range of the sodium, as indicated by the semi-variogram in Figure 16 (a), is approximately 300 m. There is a strong spatial autocorrelation in the prediction (Figure 17 (b)).

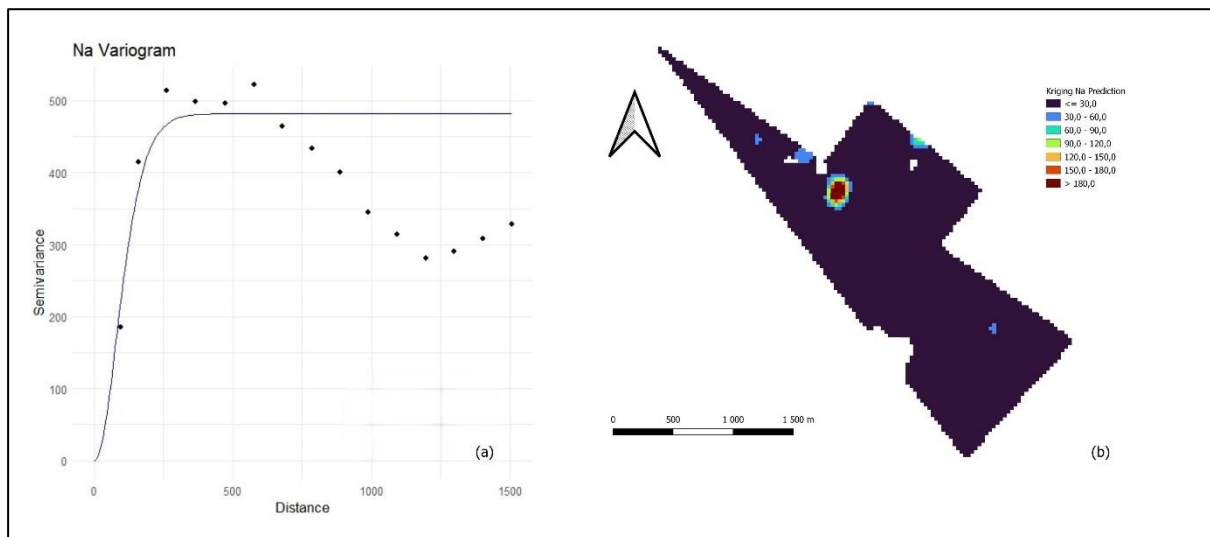


Figure 16: The constructed semi-variogram (a) and OK map (b) of the sodium (Na).

The sodium levels in the soil are low throughout the field, as shown in the sodium prediction maps in Figure 17. The higher sodium levels in the field are largely concentrated around the pan found in the north end of the field and on the edges of the field. The OK result (Figure 16 (b)) included negative sodium values, as low as -33 mg/kg. This is impossible and an obvious major flaw in the prediction. This is a known limitation of OK, as OK weights are linear and not constrained to a positive value. Predictions, therefore, can fall outside of the valid range of the original data, where further away points are assigned negative weights in order to reduce redundancy with closer points. The skewed distribution of the Na values likely increased the effect of this. The Cubist result (Figure 17 (b)) performed the best of the three methods, with the lowest RMSE, although it is still extremely high when compared to the mean of the Na values. The influence of the covariate data, particularly the satellite imagery, is apparent in the DSM method maps, and can be seen by the lines in the field that are where fence lines existed in the past. This could make sense, as liming also induces leaching of the Na ion. The Na values are high at the old fence lines and could indicate the lack of liming. The Na predictions were outperformed by those of Louw (2025), where they achieved an RMSE value of only 0.79 for OK, 1.22 for Cubist and 0.92 for RF. The CCC values they presented were, however, far lower than those found in Table 10, with -0.01, -0.11 and -0.04 for the OK, Cubist and RF respectively. Thus, the Na results of this study and those of Louw can be considered to be similar, but for the purposes of fertiliser recommendations, the result presented by Louw (2025) would be preferable.

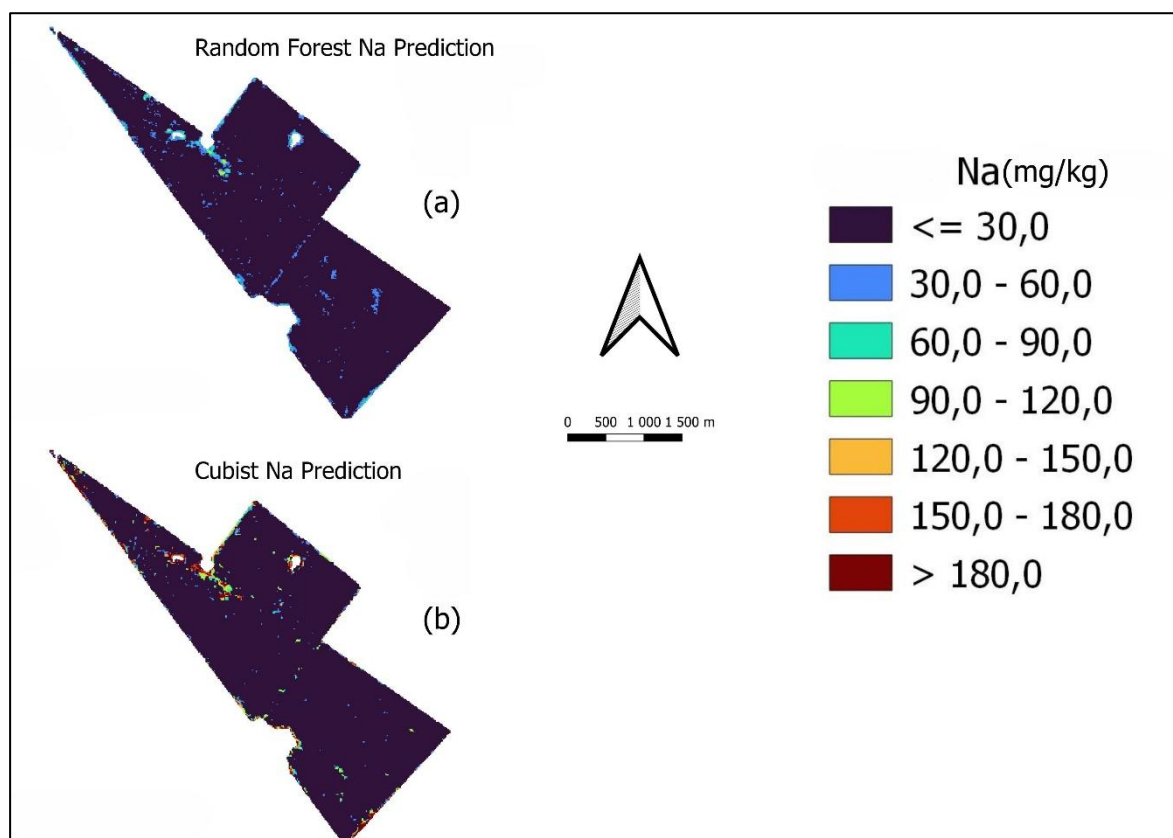


Figure 17: The generated sodium (Na) maps of RF (a) and Cubist (b).

Table 9: The accuracy statistics of the sodium (Na) maps

Na Mapping Method	RMSE (mg/kg)	CCC	MEC
OK	18.26	0.21	0.36
Cubist	6.60	0.18	-0.3
RF	7.44	0.15	-0.46

4.3.7 Cation Exchange Capacity

The range of the CEC semi-variogram (Figure 18 (a)) was found to be 3336.282 m. There does appear to be over-smoothing of the OK result, indicative of a weak spatial autocorrelation (Yamamoto, 2005). The CEC, Ca, Mg and K values all have similar trends, as all of their respective OK predictions show directionality, with similar distributions of values throughout the predictions, with high values in the southeast of the field, and lower values in the northern area. Therefore, as distance increases, variance also increases. This is thus reason for concern in industry, due to the fact that the conventional method of making OK predictions does not involve the creation of semi-variograms that the mapper examines before making the mapping prediction.

It is essential to note that this can be a problem for farmers as the spatial correlation of the points are not considered by mappers when making predictions.

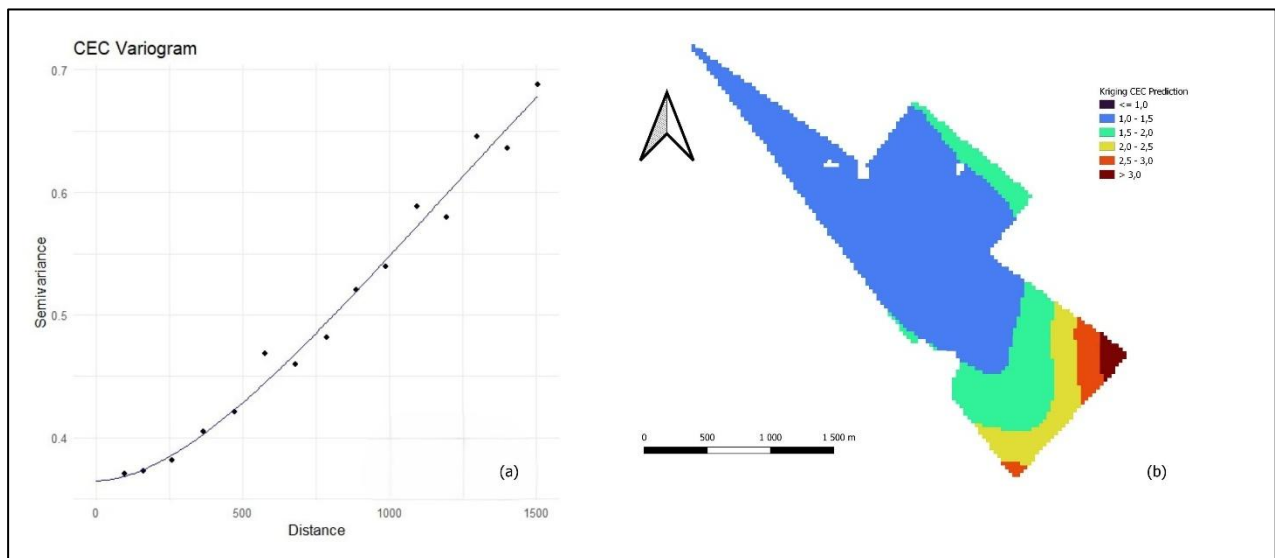


Figure 18: The constructed semi-variogram (a) and OK map (b) of the cation exchange capacity (CEC).

Figure 19 includes the DSM CEC prediction maps. The largest concentration of cations in the field lie largely in the southeast corner of the field, which is consistent with the Mg and Ca concentrations as both Mg and Ca are positively charged. All three of the maps show similar distribution, with a significant number of patchy areas in the Cubist (Figure 19 (b)) and RF (Figure 19 (a)) maps. Statistically, the Cubist and RF results were extremely comparable in this case, both outperforming the OK (Table 11).

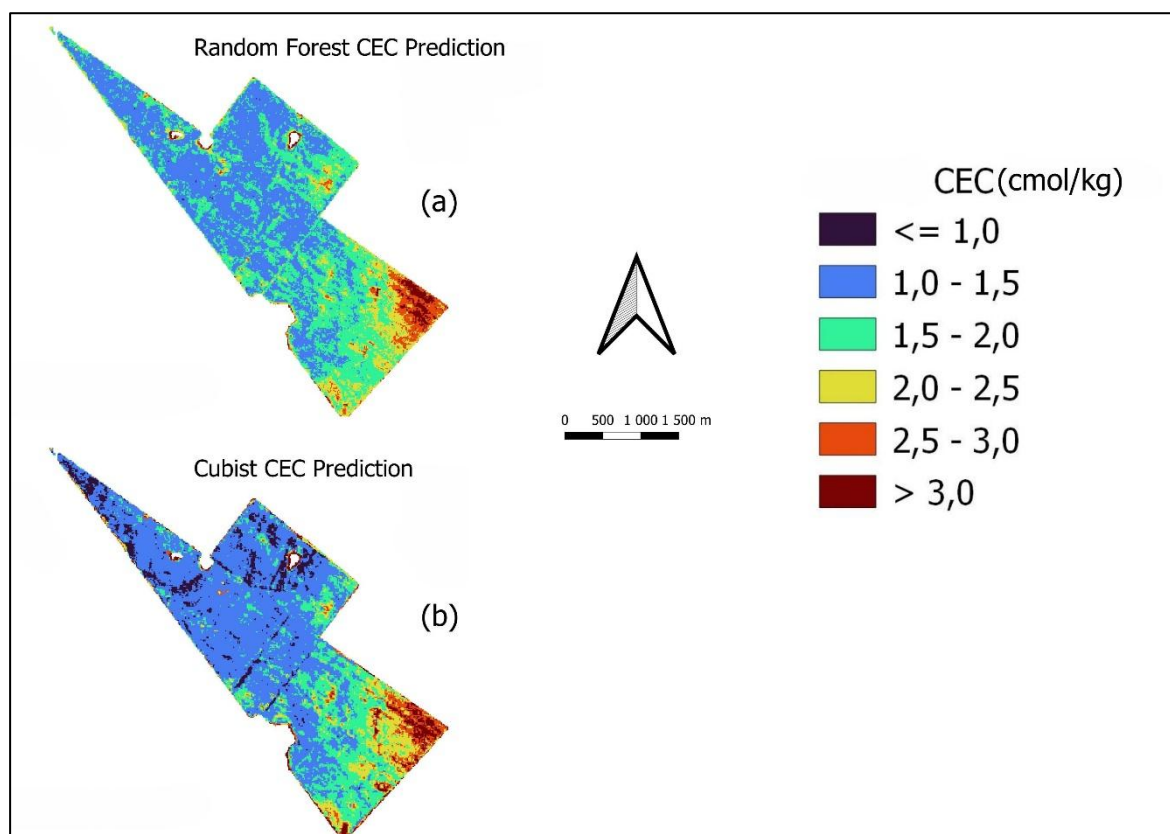


Figure 19: The generated Cation Exchange Capacity (CEC) maps of RF (a) and Cubist (b).

Table 10: The accuracy statistics of the cation exchange capacity (CEC) maps

CEC Mapping Method	RMSE (cmol(+)/kg)	CCC	MEC
OK	0.67	0.40	0.28
Cubist	0.47	0.50	0.23
RF	0.47	0.48	0.14

4.3.8 Summary

Unfortunately, model performance varied significantly throughout the properties. While certain predictions were not usable, others were more accurate and could be used. When considering the RMSE values of the P and K specifically, which are two of the important macronutrients for plant growth, the maps of all three prediction models are not usable as their RMSE values are larger than the class widths found in the Fertiliser Handbook (FERTASA, 2016). There was no clear indication that OK yielded better results than those of the ML predictions. For the conventional mapping method, the observation points were spaced too far away from one another, and the clear directionality was not considered. The DSM produced maps were not much better, even though it had very densely populated yield maps to use as covariates, which should

be correlated with the different soil properties. The poor performance of the DSM could be attributed to yield data inconsistencies, as the data was collected from two different harvesters, which perhaps were not consistently calibrated. The DSM was also limited to sampling points on the grid, and full spatial coverage of the covariate space could not be obtained. A further concern for the DSM methods is the lack of a P prediction from the Cubist model. A study by Pouladi et al. (2019) found that, when mapping soil organic matter (SOM), both kriging and regression kriging outperformed DSM using Cubist and RF. Kriging achieved an RMSE of 2.27% SOM and a MEC of 0.91, while the Cubist model achieved an RMSE of 3.26% and an MEC of 0.88, while the RF achieved a RMSE value of 4.2% and a MEC of 0.89, which are all better results than results in this study. However, the accuracy of the results in this study can be attributed to the fact that it made use of a higher density of sampling points. Furthermore, a study by Pereira et al. (2022) found that the use of OK outperformed both DSM with ML and Inverse Distance Weighted (IDW) interpolation when making predictions with a low density of samples, while the opposite was true in a situation where sample point density is low, with the DSM with ML method outperforming the OK and IDW methods. This concurs with the results of this study, even though the DSM with ML maps were still not accurate enough. Specific covariate data would improve the DSM, such as accurately calibrated yield data, and additional soil sensors such as Electromagnetic Induction (EMI) (Doolittle and Brevik, 2014) and Gamma radiometrics (Reinhardt and Herrmann, 2019).

4.4 Fertiliser Recommendation and Cost-benefit Analysis

The yield potential map of the field can be found in Figure 20. The majority of the field is comprised of areas that have a yield potential of 5 t/ha, as previously determined by the agronomist consultant assisting the farmer. This data was supplied by the farmer.

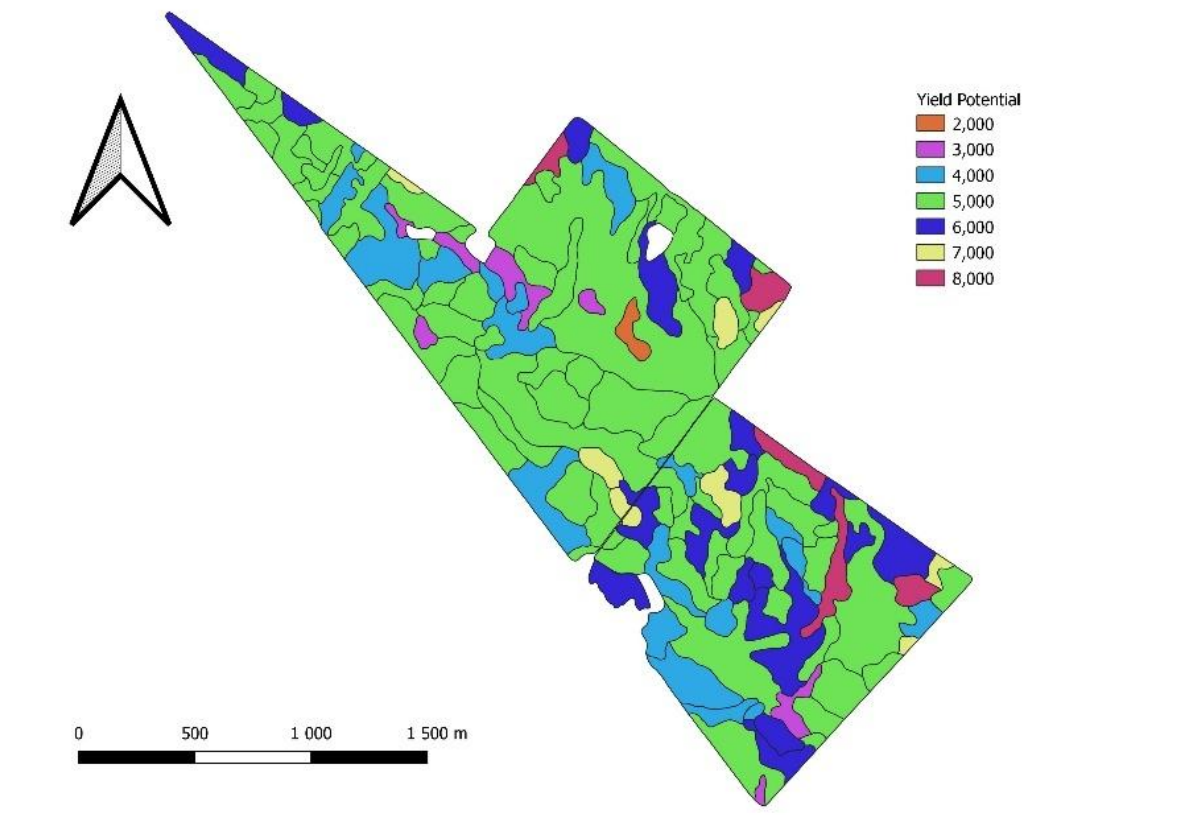


Figure 20: Yield Potential Map

4.4.1 Fertiliser Recommendation Maps

The P fertiliser recommendation maps created using the conventional method (OK) and DSM with ML can be found in Figure 21. The DSM with ML based P fertiliser recommendation maps (Figure 8) again show a lot of patchy areas, largely due to the merging of the yield potential map and the already patchy DSM with ML prediction. On the other hand, the OK based fertiliser recommendation map was far more uniform throughout the field, with only two fertiliser application classes, as compared to the three fertiliser application classes of the DSM with ML based fertiliser recommendation maps. Both results recommend 30 kg/ha needed for the majority of the field, with smaller areas of the field needing lesser amounts of P fertiliser.

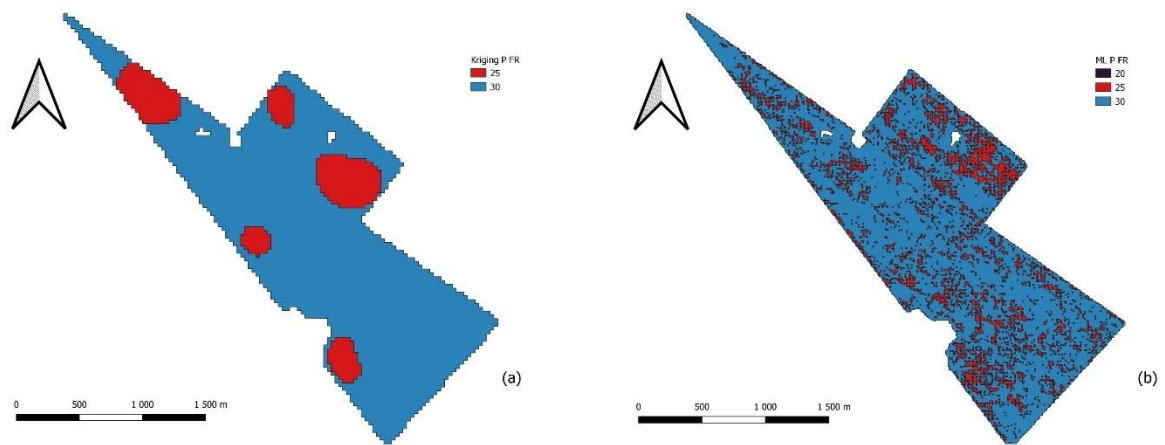


Figure 21: Map indicating the Phosphorus Fertiliser Recommendation created with conventional mapping (OK) (a) and digital soil mapping (DSM) with machine learning (ML) (b).

The K fertiliser recommendation maps are given in Figure 22. The DSM with ML based K fertiliser recommendation map (Figure 22b)) displays a large number of patchy areas throughout the field, while the OK based recommendation is far more uniform in its areas. Both maps are, however, visually similar in the majority of areas in the field. Both maps also show the need for higher K application levels in the southeast corner of the field, with less fertiliser needed through the rest of the field.

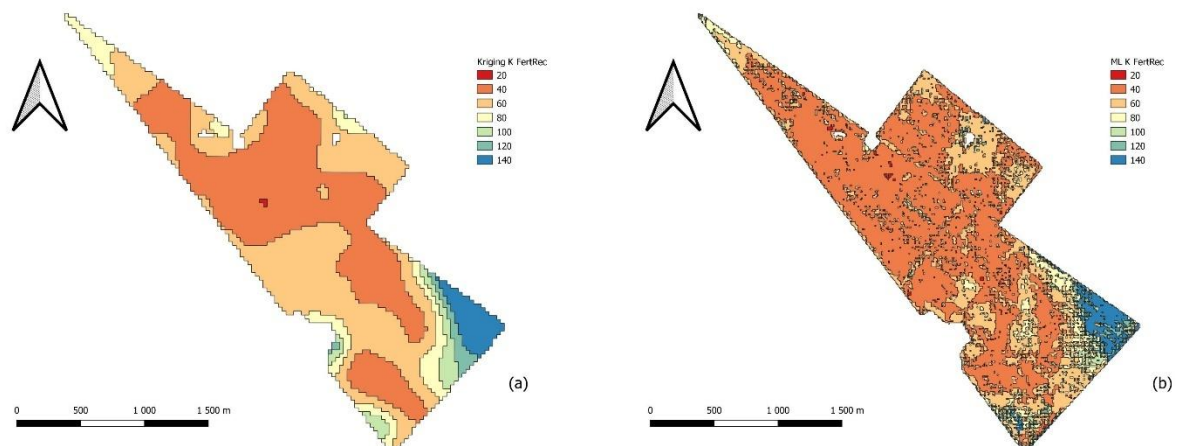


Figure 22: Maps indicating the Potassium Fertiliser Recommendation Maps created using conventional mapping (OK) (a) and digital soil mapping (DSM) with machine learning (ML) (b).

4.4.2 Financial Implications Analysis

The financial implications of the fertiliser applications can be found in Table 13. In each case, the Profit was calculated as the total income (yield × maize price), whereafter the combined fertiliser costs were subtracted (Profit = Income – (P Fertiliser Cost + K Fertiliser Cost)) The optimal scenario represents the theoretical profit that is achievable using the soil property values at the sampling points, while the DSM and OK scenarios represent the profit achievable using each

method. The optimal yield of the field was 1781 tonnes, while the DSM yield was 1328.47 tonnes. The OK yield was 1373.55 tonnes. The optimal fertiliser application had a total fertiliser cost of R809 886,44, while the DSM and OK had fertiliser costs of R677 892,60 and R644 008,32 respectively. The optimal profit was far higher than both methods, with the OK profit exceeding that of the DSM by R304 336,68. However, the difference between the optimal profit and the two methods is telling of the significance of the accuracy of the mapping methods. The loss as a result of the DSM method was R2 583 187.96, while the loss due to the OK method was R2 278 851.28. This is a massive loss of 24% and 21% respectively, no matter which method you choose, especially considering the fact that it is largely due to inaccurate soil property maps. The inaccuracy of the OK method could be due to the directionality and over smoothing of the P and K prediction maps (Yamamoto, 2005). The poor performance of the DSM with ML based fertiliser application map can largely be attributed to covariate data discrepancies, such as the harvester data being obtained from two different harvesters, as well as the inability to select sampling point locations using scientifically based methods, as well as the yield data being obtained from two different harvesters, and also the weak relationship between the covariates and soil properties.

Table 12: Fertiliser Financial Implication Values

Optimal Yield (tonnes)		1781.00
Optimal Income	R	10 686 000.00
Optimal Potassium Fertiliser Cost	R	159 138.44
Optimal Phosphorus Fertiliser Cost	R	650 748.00
Optimal Profit	R	9 876 113.56
DSM Yield (tonnes)		1328.47
DSM Income	R	7 970 818.20
DSM Potassium Fertiliser Cost	R	126 147.60
DSM Phosphorus Fertiliser Cost	R	551 745.00
DSM Profit	R	7 292 925.60
OK Yield (tonnes)		1373.55
OK Income	R	8 241 270.60
OK Potassium Fertiliser Cost	R	133 804.32
OK Phosphorus Fertiliser Cost	R	510 204.00
OK Profit	R	7 597 262.28

CHAPTER 5 CONCLUSIONS

This study aimed to determine the most accurate and cost-effective method to map soil properties at field scale. This was achieved through the completion of three main objectives. These objectives are discussed below.

The first objective was to determine whether the conventional method of soil mapping is viable and whether the distance between sampling points is correct. This objective was tested by generating semi-variograms and maps using the conventional method for several soil properties. Map accuracy for OK was evaluated using 10-fold cross-validation and the RMSE, CCC and MEC, together with visual analysis. The semi-variograms were inspected in order to determine whether the conventional, industry standard method of sampling on a 1-hectare grid and mapping soils accordingly was viable. In particular, the range of the semi-variograms was considered in order to determine the maximum distance at which samples could viably be sampled at.

The findings with regard to the first objective were that the sampling distance of 100 m associated with the 1-hectare grid system was sufficient for all of the tested properties. However, the OK predictions reflected the directional trends of the data, particularly the Na, Ca, Mg and K predictions. Furthermore, in the case of the P prediction, there was little to no spatial correlation as a result of a lack of variability in the P values in the field. The P and K predictions are completely unusable for all of the prediction models, as they all fall outside of the class widths found in the Fertiliser Handbook (FERTASA, 2016). The pH prediction had a relatively good spatial autocorrelation based on the nugget:sill ratio of the semi-variogram.

The first objective had some limitations, one being the inability to test other types of interpolation, such as IDW, regression kriging and universal kriging due to the fact that the study only tested the industry standard method, namely, OK. Furthermore, the study would have benefited from testing more than one field, with a diverse mixture of management practices, such as tillage practices and crop cultivation types.

The second objective, which was to test the viability of DSM for soil property mapping for PA, was achieved by generating maps using DSM with ML models. The models chosen were Cubist and RF. These maps were generated using only 70% of the available sampling points, while the accuracy was tested with the remaining 30%. A number of covariate data sources were utilized, such as a DEM, three years of yield data (2021, 2023 and 2024) and four satellite-derived spectral indices (BI, RI, SI and NDVI). Validation statistics used were RMSE, CCC and MEC. These statistics were examined along with the maps, and the performance of the different methods were independently evaluated.

The findings were that although the DSM produced maps showed validation statistics that were marginally favourable based on their validation approach, with the OK showing validation statistics that were under those of the DSM approaches, all of the mapping methods were too

inaccurate to be effective in industry based on their RMSE values and the class widths present in the Fertiliser Handbook (FERTASA, 2016). The OK maps had some over smoothing in the Mg prediction, while there was significant directionality in the Ca, Na and K predictions. The Cubist model could not make a prediction for P, likely due to a lack of variability in the P values in the field. One advantage of the DSM with ML is its ability to take into account factors like the historic fence lines present in the field, particularly when mapping properties that are related to liming (Ca, Mg), as the fence lines would not have been applied with lime.

The limitations of DSM were that with this study the training and validation points were limited to the grid previously sampled. This does not allow for the entire covariate space to be represented in the samples. Improved results should be gained if samples could be collected from the entire field. Additionally, specific covariate data would improve the DSM, such as accurately calibrated yield data, and additional soil sensors such as EM and Gamma radiometrics.

Kriging standard deviations could be used to construct local prediction intervals, along with bootstrapping methods can allow for calculation of the error of RF predictions. Creating and comparing 90% prediction intervals for the methods and thereafter including them into financial analysis is recommended for future work.

The third objective, which was to determine the financial implications of the two soil property mapping methods, achieved by generating fertiliser recommendation maps using the P and K predictions generated previously by the DSM with ML and OK methods, and the yield potential map of the field. Thereafter, industry pricing of both fertilisers and the crop were used to calculate the financial implications of the two mapping methods, taking into account over fertilisation, as well as the under performance of the crop due to under fertilisation. These results were then compared in order to determine the loss of profit due to inaccuracy of the mapping methods.

The findings of the financial analysis were that profit loss is directly linked to the accuracy of the map, which will negatively impact the fertiliser applications if it is inaccurate. In order for predictions to be accurate enough, the RMSE of the predictions should fall within the class sizes found in the Fertiliser Handbook (FERTASA, 2016). Both DSM and OK lead to huge loss of revenue, largely due to under fertilisation. The losses of the DSM method were R2 583 187.96, while the OK method had losses of R2 278 851.28. While the OK method lost less than the DSM method, this is still a major cost implication due to inaccurate mapping methods and therefore, it should be addressed. The likely reason that OK had a lower financial loss than that of the DSM methods is that the prediction was constrained by the variogram and smoothed the prediction. This produced a more conservative fertiliser recommendation than that of the DSM, which produced more extreme predictions due to the covariate-dependent approach of the method.

Some of the limitations of the study were the inability to select data points using better statistical methods. Furthermore, yield data from the same harvester or harvesters that are calibrated in the

same way would be beneficial in order to avoid variation in the data. A lack of variance in the P data from the field also limited the Cubist model from creating a P prediction.

Future research should be conducted with more interpolation methods such as regression kriging, in order to better understand the abilities of each of the methods and address the directionality displayed in this study. A weak relationship between the covariates and soil properties should be the focus of future study, with a main focus on expanding the number of covariates into different types of covariates such as electromagnetic induction or gamma-ray spectroscopy. A better data outlay could also be beneficial and used to determine the correct grid size in order to better guide the agricultural industry. Furthermore, similar research should be conducted on a number of different fields with differing management practices, to better understand the performance of the methods when faced with different situations. Improved covariates will also improve the accuracy of the DSM predictions.

REFERENCES

- Adeniyi, O.D., Bature, H. and Mearker, M. (2024) "A systematic review on digital soil mapping approaches in lowland areas," *Land*, 13(3), p. 379. DOI: <https://doi.org/10.3390/land13030379>.
- Ahuja, S. and Angra, S. (2017) *Machine learning and its Applications: A Review*, *Researchgate.net*. DOI: <https://doi.org/10.1109/ICBDACI.2017.8070809>.
- Arrouays, D. et al. (2020) "Impressions of digital soil maps: The good, the not so good, and making them ever better," *Geoderma regional*, 20(e00255), p. e00255. DOI: <https://doi.org/10.1016/j.geodrs.2020.e00255>.
- Arrouays, D., Lagacherie, P. and Hartemink, A.E. (2017) "Digital soil mapping across the globe," *Geoderma regional*, 9, pp. 1–4. DOI: <https://doi.org/10.1016/j.geodrs.2017.03.002>.
- Berger, T.W. (2020) "Challenges and lessons for measuring soil metrics in household surveys," *Soil Science Society of America Journal* [Preprint].
- semi-variogramda Silveira, F., Lermen, F.H. and Amaral, F.G. (2021) "An overview of agriculture 4.0 development: Systematic review of descriptions, technologies, barriers, advantages, and disadvantages," *Computers and electronics in agriculture*, 189(106405), p. 106405. DOI: <https://doi.org/10.1016/j.compag.2021.106405>.
- Demir, S. and Sahin, E.K. (2024) "The effectiveness of data pre-processing methods on the performance of machine learning techniques using RF, SVR, Cubist and SGB: a study on undrained shear strength prediction," *Stochastic environmental research and risk assessment: research journal*, 38(8), pp. 3273–3290. DOI: <https://doi.org/10.1007/s00477-024-02745-9>.
- Dharumarajan, S. and Hegde, R. (2022) "Digital mapping of soil texture classes using Random Forest classification algorithm," *Soil use and management*, 38(1), pp. 135–149. DOI: <https://doi.org/10.1111/sum.12668>.
- Doolittle, J.A. and Brevik, E.C. (2014) "The use of electromagnetic induction techniques in soils studies," *Geoderma*, 223–225, pp. 33–45. DOI: <https://doi.org/10.1016/j.geoderma.2014.01.027>.
- du Plessis, C. et al. (2020) "Machine learning digital soil mapping to inform gully erosion mitigation measures in the Eastern Cape, South Africa," *Geoderma*, 368(114287), p. 114287. DOI: <https://doi.org/10.1016/j.geoderma.2020.114287>.

- Earth Explorer. 2024. Sentinel 2 Satellite Imagery. Date of Access: 13 Nov. 2024
- Environmental Systems Research Institute, Inc. (ESRI). 2020. Generalization toolset: Union
- Erickson, B. and Fausti, S.W. (2021) "The role of precision agriculture in food security," *Agronomy journal*, 113(6), pp. 4455–4462. DOI: <https://doi.org/10.1002/agj2.20919>.
- Esfandiarpour Borujeni, I. *et al.* (2010) "Assessing geopedological soil mapping approach by statistical and geostatistical methods: A case study in the Borujen region, Central Iran," *Catena*, 82(1), pp. 1–14. DOI: <https://doi.org/10.1016/j.catena.2010.03.006>.
- Fernández-Delgado, M. *et al.* (2019) "An extensive experimental survey of regression methods," *Neural networks: the official journal of the International Neural Network Society*, 111, pp. 11–34. DOI: <https://doi.org/10.1016/j.neunet.2018.12.010>.
- FERTASA (2016) *FSSA Fertiliser handbook of South Africa*. 7th ed. The Fertiliser Society of Southern Africa.
- Fertiliser Association of South Africa (FERTASA) (2016) *Fertiliser Handbook*. 7th Revised Edition. Edited by E. Greenfield *et al.* Lynnwood Ridge: Fertiliser Association of South Africa (FERTASA).
- Gebbers, R. and Adamchuk, V.I. (2010) "Precision agriculture and food security," *Science (New York, N.Y.)*, 327(5967), pp. 828–831. DOI: <https://doi.org/10.1126/science.1183899>.
- Gong, Z. *et al.* (2003) "Origin and development of soil science in ancient China," *Geoderma*, 115(1–2), pp. 3–13. DOI: [https://doi.org/10.1016/s0016-7061\(03\)00071-5](https://doi.org/10.1016/s0016-7061(03)00071-5).
- Grafton, R.Q., Daugbjerg, C. and Qureshi, M.E. (2015) "Towards food security by 2050," *Food security*, 7(2), pp. 179–183. DOI: <https://doi.org/10.1007/s12571-015-0445-x>.
- Grimm, R. *et al.* (2008) "Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis," *Geoderma*, 146(1–2), pp. 102–113. DOI: <https://doi.org/10.1016/j.geoderma.2008.05.008>.
- Hateffard, F., Steinbuch, L. and Heuvelink, G.B.M. (2024) "Evaluating the extrapolation potential of random forest digital soil mapping," *Geoderma*, 441(116740), p. 116740. DOI: <https://doi.org/10.1016/j.geoderma.2023.116740>.

- Fichtner, T., Goersmeyer, N. and Stefan, C. (2019) "Influence of soil pore system properties on the degradation rates of organic substances during soil aquifer treatment (SAT)," *Applied Sciences* (Basel, Switzerland), 9(3), p. 496. DOI: <https://doi.org/10.3390/app9030496>
- Heil, J., Jörges, C. and Stumpe, B. (2022) "Fine-scale mapping of soil organic matter in agricultural soils using UAVs and machine learning," *Remote sensing*, 14(14), p. 3349. DOI: <https://doi.org/10.3390/rs14143349>.
- Hempel, J.W. *et al.* (2008) *Digital Soil Mapping with Limited Data*. Edited by A.E. Hartemink. New York, NY: Springer Science+Business Media.
- Hodson, T.O. (2022) "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geoscientific model development*, 15(14), pp. 5481–5487. DOI: <https://doi.org/10.5194/gmd-15-5481-2022>.
- Huang, H. *et al.* (2022) "A review on digital mapping of soil carbon in cropland: progress, challenge, and prospect," *Environmental research letters*, 17(12), p. 123004. DOI: <https://doi.org/10.1088/1748-9326/aca41e>.
- Jacobs, A.J., van Tol, J.J. and du Preez, C.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 (SAJAE)*, 46(2), pp. 107–118. DOI: <https://doi.org/10.17159/2413-3221/2018/v46n2a484>.
- Kaya, F. *et al.* (2022) "Assessing machine learning-based prediction under different agricultural practices for digital mapping of soil organic carbon and available phosphorus," *Agriculture*, 12(7), p. 1062. DOI: <https://doi.org/10.3390/agriculture12071062>.
- Khaledian, Y. and Miller, B.A. (2020) "Selecting appropriate machine learning methods for digital soil mapping," *Applied mathematical modelling*, 81, pp. 401–418. DOI: <https://doi.org/10.1016/j.apm.2019.12.016>.
- Kidd, D. *et al.* (2020) "Operationalising digital soil mapping – Lessons from Australia," *Geoderma regional*, 23(e00335), p. e00335. DOI: <https://doi.org/10.1016/j.geodrs.2020.e00335>.
- Kuhn, M. and Quinlan, R. (2011) "Cubist: Rule- and instance-based regression modeling," *CRAN: Contributed Packages*. The R Foundation.
- Kuhn, M. *et al.* (2023) *Package 'Cubist,' R-project.org*. DOI: <https://cran.r-project.org/web/packages/Cubist/Cubist.pdf> (Accessed: March 18, 2024).

- Lamichhane, S., Kumar, L. and Wilson, B. (2019) "Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review," *Geoderma*, 352, pp. 395–413. DOI: <https://doi.org/10.1016/j.geoderma.2019.05.031>.
- Land Type Survey Staff (1972) Land types of South Africa: Digital map (1:250 000 scale) and soil inventory datasets. Pretoria: Agricultural Research Council – Institute for Soil, Climate and Water.
- Lin, L.I. (1989) "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, 45(1), pp. 255–268. DOI: <https://doi.org/10.2307/2532051>.
- Loiseau, T. *et al.* (2021) "Density of soil observations in digital soil mapping: A study in the Mayenne region, France," *Geoderma regional*, 24(e00358), p. e00358. DOI: <https://doi.org/10.1016/j.geodrs.2021.e00358>.
- Louw, E.H. (2025) "Applying machine learning digital soil mapping techniques on farm scale for use in precision agriculture," Unpublished Masters Dissertation.
- Lowenberg-DeBoer, J. and Erickson, B. (2019) "Setting the record straight on precision agriculture adoption," *Agronomy journal*, 111(4), pp. 1552–1569. DOI: <https://doi.org/10.2134/agronj2018.12.0779>.
- Mahesh, B. (2019) "Machine Learning Algorithms - A Review," *International Journal of Science and Research (IJSR)*, 9(1), pp. 380–386.
- Malone, B.P., Minasny, B. and McBratney, A.B. (2016) *Using R for digital soil mapping*. 1st ed. Cham, Switzerland: Springer International Publishing.
- management/dissolve.htm Date of access: 14 Nov. 2024
- Mazzoncini, M. (2024) "An improved digital soil mapping approach to predict total N by combining machine learning algorithms and open environmental data," *Modeling Earth Systems and Environment* [Preprint].
- McBratney, A.B., Mendonça Santos, M.L. and Minasny, B. (2003) "On digital soil mapping," *Geoderma*, 117(1–2), pp. 3–52. DOI: [https://doi.org/10.1016/s0016-7061\(03\)00223-4](https://doi.org/10.1016/s0016-7061(03)00223-4).
- Meng, Q., Liu, Z. and Borders, B.E. (2013) "Assessment of regression kriging for spatial interpolation – comparisons of seven GIS interpolation methods," *Cartography and geographic information science*, 40(1), pp. 28–39. DOI: <https://doi.org/10.1080/15230406.2013.762138>.

- Minasny, B. and McBratney, A.B. (2006) "A conditioned Latin hypercube method for sampling in the presence of ancillary information," *Computers & geosciences*, 32(9), pp. 1378–1388. DOI: <https://doi.org/10.1016/j.cageo.2005.12.009>.
- Minasny, B. and McBratney, A.B. (2016) "Digital soil mapping: A brief history and some lessons," *Geoderma*, 264, pp. 301–311. DOI: <https://doi.org/10.1016/j.geoderma.2015.07.017>.
- Monteiro, A., Santos, S. and Gonçalves, P. (2021) "Precision agriculture for crop and livestock farming-brief review," *Animals: an open access journal from MDPI*, 11(8), p. 2345. DOI: <https://doi.org/10.3390/ani11082345>.
- Munyati, C. and Sinthumule, N.I. (2021) "Comparative suitability of ordinary kriging and Inverse Distance Weighted interpolation for indicating intactness gradients on threatened savannah woodland and forest stands," *Environmental and Sustainability Indicators*, 12(100151), p. 100151. DOI: <https://doi.org/10.1016/j.indic.2021.100151>.
- Nasteski, V. (2017) "An overview of the supervised machine learning methods," *HORIZONS.B*, 4, pp. 51–62. DOI: <https://doi.org/10.20544/horizons.b.04.1.17.p05>.
- Nguyen, H.T. (2019) "Application of Ordinary Kriging and Regression Kriging Method for Soil Properties Mapping in Hilly Region of Central Vietnam," *ISPRS International Journal of Geo-Information* [Preprint].
- Noi, P., Degener, J. and Kappas, M. (2017) "Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data," *Remote sensing*, 9(5), p. 398. DOI: <https://doi.org/10.3390/rs9050398>.
- O'Shaughnessy, S.A., Evett, S.R. and Colaizzi, P.D. (2015) "Dynamic prescription maps for site-specific variable rate irrigation of cotton," *Agricultural water management*, 159, pp. 123–138. DOI: <https://doi.org/10.1016/j.agwat.2015.06.001>.
- Webster, R. and Oliver, M.A. (2007) *Geostatistics for Environmental Scientists*. 2nd ed. Hoboken, NJ: Wiley-Blackwell.
- Omer, B. et al. (2024) "Toward precision agriculture: Integrating machine learning techniques for smart farming systems," *IEEE access: practical innovations, open solutions*, 12, pp. 172910–172922. DOI: <https://doi.org/10.1109/access.2024.3480868>.

- Onyango, C.M. *et al.* (2021) "Precision agriculture for resource use efficiency in smallholder farming systems in Sub-Saharan Africa: A systematic review," *Sustainability*, 13(3), p. 1158. DOI: <https://doi.org/10.3390/su13031158>.
- Pebesma, E.J. (2004) "Multivariable geostatistics in S: the gstat package," *Computers & geosciences*, 30(7), pp. 683–691. DOI: <https://doi.org/10.1016/j.cageo.2004.03.012>.
- Pereira, G.W. *et al.* (2022) "Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting," *Precision agriculture*, 23(4), pp. 1189–1204. DOI: <https://doi.org/10.1007/s11119-022-09880-9>.
- Phan, N. (2023) *Computational method of ArcGIS Ordinary Kriging tool*, Esri.ca. DOI: <https://ecce.esri.ca/unb-blog/2023/01/20/arcgis-ordinary-kriging-tool/> (Accessed: March 4, 2025).
- Pouladi, N. *et al.* (2019) "Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging," *Geoderma*, 342, pp. 85–92. DOI: <https://doi.org/10.1016/j.geoderma.2019.02.019>.
- QGIS Development Team. (2023). *QGIS Geographic Information System, Version 3.28*. Open Source Geospatial Foundation Project. DOI: <http://qgis.org>.
- R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria.
- Reinhardt, N. and Herrmann, L. (2019) "Gamma-ray spectrometry as versatile tool in soil science: A critical review," *Journal of plant nutrition and soil science*, 182(1), pp. 9–27. DOI: <https://doi.org/10.1002/jpln.201700447>.
- Robinson, T.P. and Metternicht, G. (2006) "Testing the performance of spatial interpolation techniques for mapping soil properties," *Computers and electronics in agriculture*, 50(2), pp. 97–108. DOI: <https://doi.org/10.1016/j.compag.2005.07.003>.
- Ryu, J.-S. *et al.* (2002) "Kriging interpolation methods in geostatistics and DACE model," *KSME international journal*, 16(5), pp. 619–632. DOI: <https://doi.org/10.1007/bf03184811>.
- Sarker, I.H. (2021) "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, 2(3). DOI: <https://doi.org/10.1007/s42979-021-00592-x>.

- Say, S.M. *et al.* (2018) "Adoption of Precision Agriculture Technologies in Developed and Developing Countries," *The Online Journal of Science and Technology*, 8(1), pp. 7–15.
- Schulze, R.E. (2007) "South African Atlas of Climatology and Agrohydrology," *Water Sinergise Solutions d. o. o.*, a Planet Labs company (2024) *Sentinel Hub*. DOI: <https://www.sentinel-hub.com/> (Accessed: 2024).
- Stein, M.L. (1999) *Interpolation of spatial data: Some theory for kriging*. 1999th ed. New York, NY: Springer.
- United Nations (2023) *The 17 Goals*, *Sdgs.un.org*. DOI: <https://sdgs.un.org/goals> (Accessed: March 7, 2024).
- van Biljon, J.J., Fouche, D.S. and Botha, A.D.P. (2008) "Threshold values and sufficiency levels for potassium in maize producing sandy soils of South Africa," *Suid-Afrikaanse tydskrif vir plant en grond [South African journal of plant and soil]*, 25(2), pp. 65–70. DOI: <https://doi.org/10.1080/02571862.2008.10639897>.
- van der Westhuizen, S., Heuvelink, G.B.M. and Hofmeyr, D.P. (2023) "Multivariate random forest for digital soil mapping," *Geoderma*, 431(116365), p. 116365. DOI: <https://doi.org/10.1016/j.geoderma.2023.116365>.
- Venter, Z.S. *et al.* (2021) "Mapping soil organic carbon stocks and trends with satellite-driven high resolution maps over South Africa," *The Science of the total environment*, 771(145384), p. 145384. DOI: <https://doi.org/10.1016/j.scitotenv.2021.145384>.
- Verster, E., du Plessis, M.J. and Schoeman, J.L. (2022) *Guidelines for Conducting Soil Surveys in South Africa*. South Africa: Soil Science Society of South Africa.
- Wadoux, A.M.J.-C. *et al.* (2021) "Spatial cross-validation is not the right way to evaluate map accuracy," *Ecological modelling*, 457(109692), p. 109692. DOI: <https://doi.org/10.1016/j.ecolmodel.2021.109692>.
- Wadoux, A.M.J.-C., Brus, D.J. and Heuvelink, G.B.M. (2019) "Sampling design optimization for soil mapping with random forest," *Geoderma*, 355(113913), p. 113913. DOI: <https://doi.org/10.1016/j.geoderma.2019.113913>.
- Wadoux, A.M.J.-C., Minasny, B. and McBratney, A.B. (2020) "Machine learning for digital soil mapping: Applications, challenges and suggested solutions," *Earth-science reviews*, 210 (103359), p. 103359. DOI: <https://doi.org/10.1016/j.earscirev.2020.103359>.

- Wadoux, A.M.J.-C., Walvoort, D.J.J. and Brus, D.J. (2022) "An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams," *Geoderma*, 405(115332), p. 115332. DOI: <https://doi.org/10.1016/j.geoderma.2021.115332>.
- Whelan, B. and Taylor, J. (2010) "Module A - Advanced," in *PA Education and Training Modules for the Grains Industry*. Sydney, Australia: Australian Centre for Precision Agriculture, University of Sydney for the Grains Research and Development Corporation.
- Yamamoto, J.K. (2005) "Correcting the smoothing effect of ordinary kriging estimates," *Mathematical geology*, 37(1), pp. 69–94. DOI: <https://doi.org/10.1007/s11004-005-8748-7>.
- DOI:Zhou, J. *et al.* (2019) "Random forests and Cubist algorithms for predicting shear strengths of rockfill materials," *Applied sciences (Basel, Switzerland)*, 9(8), p. 1621. DOI: <https://doi.org/10.3390/app9081621>.
- Zhu, Q. and Lin, H.S. (2010) "Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes," *Pedosphere*, 20(5), pp. 594–606. DOI: [https://doi.org/10.1016/s1002-0160\(10\)60049-5](https://doi.org/10.1016/s1002-0160(10)60049-5).

