Predicting spatiotemporal yield variability to aid arable precision agriculture in New Zealand: A case study of maize-grain crop production in the Waikato Region

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Agriculture and Horticulture

At

Massey University

Palmerston North

New Zealand

Guopeng Jiang

2020

Thesis abstract

Precision agriculture attempts to manage within-field spatial variability by applying suitable inputs at the appropriate time, place, and amount. To achieve this, delineation of field-specific management zones (MZs), representing significantly different yield potentials are required. To date, the effectiveness of utilising MZs in New Zealand has potentially been limited due to a lack of emphasis on the interactions between spatiotemporal factors such as soil texture, crop yield, and rainfall. To fill this research gap, this thesis aims to improve the process of delineating MZs by modelling spatiotemporal interactions between spatial crop yield and other complementary factors.

Data was collected from five non-irrigated field sites in the Waikato region, based on the availability of several years of maize harvest data. To remove potential yield measurement errors and improve the accuracy of spatial interpolation for yield mapping, a customised filtering algorithm was developed. A supervised machine-learning approach for predicting spatial yield was then developed using several prediction models (stepwise multiple linear regression, feedforward neural network, CART decision tree, random forest, Cubist regression, and XGBoost). To provide insights into managing spatiotemporal yield variability, predictor importance analysis was conducted to identify important yield predictors.

The spatial filtering method reduced the root mean squared errors of kriging interpolation for all available years (2014, 2015, 2017 and 2018) in a tested site, suggesting that the method developed in R programme was effective for improving the accuracy of the yield maps. For predicting spatial yield, random forest produced the highest prediction accuracies (R^2 = 0.08 - 0.50), followed by XGBoost (R^2 = 0.06 - 0.39). Temporal variables (solar radiation, growing degree days (GDD) and rainfall) were proven to be salient yield predictors. This research demonstrates the viability of these models to predict subfield spatial yield, using input data that is inexpensive and readily available to arable farms in New Zealand. The novel approach employed by this thesis may provide opportunities to improve arable farming input-use efficiency and reduce its environmental impact.

Acknowledgement

I would like to express my gratitude to New Zealand Foundation for Arable Research for providing the scholarship and data for this PhD project. I would like to thank Allister Holmes (FAR Research & Extension Team Leader) for mentoring me throughout this project. Allister, a man of great character, has always been positive and supportive of my work. I would like to thank Professor Peter Kemp (Head of School of Agriculture and Environment, Massey University) for granting some extra funding. This project could not have been completed without the financial support from the college.

I would like to thank my supervisors: Dr Miles Grafton, Professor Diane Pearson and Dr Mike Bretherton for their academic guidance in the construction of this thesis. I am grateful for Miles for his trust, expert advice and great patience during the iterations of this thesis. Professor Diane Pearson, I am privileged to have had the opportunity to be your student and truly appreciate your constructive feedback and insightful comments which were essential to consolidating the final work. I would like to thank Dr Mike Bretherton, a trained soil scientist for the quality advice and for helping me with soil coring, preparation of soil experiments, and reviewing my thesis and papers for publication. I would also like to thank Dr Anja Mobis and other laboratory staff for the instruction in the lab.

Most importantly, I am grateful for my family back in Guangzhou, China. My grandfather (deceased) and grandmother especially, for their unfading love and inspiration that gives me the power to take on this challenging task. My grandmother, a retired high-school teacher, plays a key role in my education and my decision to pursue a PhD degree. For this, I dedicate this thesis to my grandfather and grandmother for giving me the best of everything throughout my life.

Contents

List of Abbreviations

Abbreviation	Explanation
CART	Classification and regression tree
CRZ	Crop reflectance zones
EC	Electrical conductivity
FAO	Food and Agriculture Organization of the United Nations
FAR	Foundation for Arable Research
FFNN	Feedforward backpropagation neural network
GDD	Growing degree days
ML	Machine learning
MPI	Ministry for Primary Industries
MZ	Management zones
NCRS	Northern Crop Research Site
NIWA	National Institute of Water and Atmospheric Research
NZ	New Zealand
OM	Organic matter
PCA	Principal component analysis
RF	Random forest
RMSE	Root mean square error
SMLR	Stepwise multiple linear regression
SSCM	Site-specific crop management
SZ	Soil zones
VRS	Variable-rate seeding
XGBoost	Extreme gradient boosting
YPZ	Yield productivity zones

List of Figures

List of Tables

Chapter 1 Introduction, aims, and objectives

1.1 Project support and overview

The Foundation for Arable Research (FAR) is the sponsor of this PhD research project undertaken at Massey University, as a part of the Sustainable Farming Fund Project No. 407932 "Transforming Variability to Profitability" funded by the New Zealand's Ministry for Primary Industries (MPI). Formed in 1995, FAR is an applied research organisation responsible to New Zealand arable growers. FAR collects an Arable Commodity Levy at the first point of sale for all cereal grain, and on maize seed, and then invests in research and technology transfer for the further development of the industry. The project was undertaken in conjunction with FAR to provide a method that would be suitable for utilising historical yield monitor data to help enable uptake of precision farming practices in New Zealand crop production. FAR provided the data and access to their research sites for conducting this study.

1.2 Focus of study

1.2.1 Small-scale production and production costs

Arable production is an essential part of the New Zealand farming system, serving mainly the domestic market for compound livestock feed production (pellets). According to Statistics NZ (2018b), most of the arable land is planted in cereals (448,777 ha in total, which is 1.7% of the total farmed area in New Zealand), with over half a million tonnes of cereal products used by the pork, poultry, beef and dairy industries. Canterbury is the largest arable producing region, with approximately 52% of the arable land in the country, producing mainly wheat, barley and peas. Maize-grain crops are mostly grown in the North Island around the Waikato, Hawke's Bay, Manawatu-Wanganui, Gisborne and Bay of Plenty, with a total of 180,000 tonnes produced and 16,000 ha grown each year (Statistics NZ, 2018b). Driven by the growth of the domestic and international economy, in 2023, the arable sector is expected to grow at around 2% in export revenue per annum (MPI, 2018).

Given a domestic focus, New Zealand arable production typically operates on a small scale. For example, an average Australian arable farmer plants around 500 ha wheat, whereas a New Zealand farmer typically has less than 100 ha (Millner et al., 2013). The milling and the feed industry must rely on imports due to high domestic transport costs (Millner et al., 2013). On the individual field level, some crops such as maize are often produced from small blocks with an average size of 8 ha and then rotated with a forage crop or annual pasture during winter for livestock grazing (Beef + Lamb New Zealand, 2019).

The production of crops in small fields also constrains working efficiency. In some developed countries, small fields are merged into a large unit, which then allows larger machines to be used. In the US, a maize planter with over 16 rows is standard, whereas, in New Zealand, the fields are typically planted with 6-8 rows planters. In this case, a US farmer can finish planting faster with less time turning around at the field boundaries, which may require extra fuel and labour for the same operation in New Zealand. Meanwhile, field merging exposes the effect of the underlying soil variability on yield, which provides a strong proposition for large crop farms to adopt variable rate technologies on planters to allow them to switch rates automatically across different soil types (Velandia et al., 2013). Also, a farm with large cropping areas over the years is more likely to receive benefits from undertaking variable rate applications than small blocks, as the costs of investment are spread over a large area to achieve a lower "dollar per hectare" cost (Godwin et al., 2003). Empirically, this is generally not the case for New Zealand due to small-scale production.

Based on an economic modelling analysis by MPI (2012) between 2010-2012 for a typical Canterbury arable farm of mixed livestock and cropping activities, it was estimated there was a major increase of in the cost of fertiliser ($369 to $421 per ha) and electricity for powering the irrigation ($60 to $82 per ha), while other working costs remain relatively stable over the period. Given the uncertainty in returns, to maximise the profitability of crop production the industry must pursue higher input-use efficiencies, through adopting new technologies and practices from international counterparts and experimenting with different production systems. The precision farming philosophy, which emphasises managing spatial

field variability by applying the appropriate inputs “at the right amount, at the right place and the right time”, may provide an opportunity to improve the sustainability and profitability of arable production in New Zealand.

1.2.2 Highly variable weather pattern

As a country comprised of two long and narrow main islands, New Zealand has a unique temperate marine climate due to its location in the southwestern Pacific Ocean. New Zealand experiences evenly dispersed annual rainfall and mild annual temperature fluctuation, which provides the opportunity to grow a variety of arable crops successfully (NIWA, 2016).

However, within seasonal spectrums, the weather pattern of New Zealand is highly variable compared to some continental climates, which tend to have more stable seasonal patterns or largely flat geological landscapes. Spatially, weather conditions vary between regions from extremely wet on the West Coast of the South Island (> 3000 mm median annual rainfall) to almost semi-arid in Central Otago and the Mackenzie Basin of inland Canterbury (700-800 mm median annual rainfall) (NIWA, 2016). To the west, the major crop production region of Canterbury can experience dry spells in summer, which makes crop production difficult without irrigation (Saunders & Saunders, 2012).

Generally, more rain falls in winter than in summer for most regions of the country. However, given the combined effect of the sea, wind and mountains, some crop-growing regions can experience dramatic weather changes (e.g. rain and temperature) within days or even hours. These large temporal variations influence many aspects of crop management such as the timing of planting and fertiliser application.

Therefore, to determine if the philosophy of precision farming applies to New Zealand crop production, there must be a consideration of local climate and weather conditions.

1.3 Need for research

1.3.1 Slow uptake of precision farming practices

The adoption of precision farming technologies started in the early 1990s with wide-spread adoption of geospatial yield monitoring systems on grain combine harvesters. However, there has been a persisting challenge in integrating yield maps into making field crop management decisions. In the US, Schimmelpfennig (2016) reported that yield monitors were the most widely adopted technology, with about half of all maize and soybean farms having yield monitors between 2010 and 2012; especially in large maize farms (70-80%). For all maize and soybean farms that adopted yield monitors, only about 25% performed actual yield mapping due to a lack of supporting services and agronomic guidance to interpret yield maps and make recommendations based on the maps (Griffin et al., 2008).

There is no specific data on the uptake of the yield mapping practice in the New Zealand arable industry, but empirically, it is likely to be lower than that in the US due to smaller-scale production. With some automated technologies, such as GPS auto-steer on tractors and boom control systems on centre pivot irrigators, the uptake has been limited to large arable farms. Small arable growers often rely on contractors to uptake technologies when they renew their equipment and have less or no incentive to purchase expensive equipment. In comparison, yield monitors are common in most combine harvesters as they are mostly factory-fitted. However, many New Zealand farmers do not calibrate yield monitors for collecting accurate data, and some do not even transfer data to an office computer, which is a similar situation to that seen in the US (Griffin et al., 2007).

Growers have lacked a clear purpose for collecting yield monitor data, apart from monitoring crop moisture (80% of the total responded population to the USDA-ARMS survey), documenting yields (50%) and helping field trials (40%) (Griffin et al., 2007). From the research level, yield maps are often used for delineating management zones (Blackmore et al., 2003; Holmes & Jiang, 2018). However, yield maps appear to have limited value on their own because they provide no identification of why the yield had varied. With the advent of field sensing technologies, there is an opportunity to extract useful information

from various spatiotemporal factors such as soil, plant, and weather, which may help to improve the input-use efficiencies for a small-scale crop production system.

1.3.2 Research trend in precision farming

There has been considerable precision farming research globally, indicated by the number of papers presented at the International Conference on Precision Agriculture (ICPA) since the 1990s. The number of papers increased from only 43 papers presented at 1st ICPA (Minneapolis, USA) in 1992 to 233 papers presented at 5th ICPA in 2000. The number of participants increased from 173 in 1992 to about 700 in 2000, with the number of participating countries increasing from 6 to around 30 over this time (Tremblay, 2015). In 2018, over 400 abstracts were submitted and reviewed for the 14th ICPA from over 30 countries (Rund, 2018). Cushnahan et al. (2017) investigated a large dataset of publications such as Elsevier's Scopus, Google Scholar and Web of Science over the entire Precision Agriculture Journal catalogue and papers presented at the International Conference for Precision Agriculture (1998, 2000, 2008, 2010, 2012, 2014, 2016), Proceedings of First Workshop in Soil Specific Crop Management (1993) and European Precision Agriculture Conferences. They found that the number of scientific publications related to data-intensive technologies such as spectral sensors has dramatically increased (Figure 1-1). Despite the advance of spatial sensors, there is a lack of focus on temporal variations (McBratney et al., 2005). Practically, it is often difficult to prescribe crop management inputs using sensor-derived information without an understanding of potential constraints imposed by temporal factors such as rainfall and solar radiation on yield.

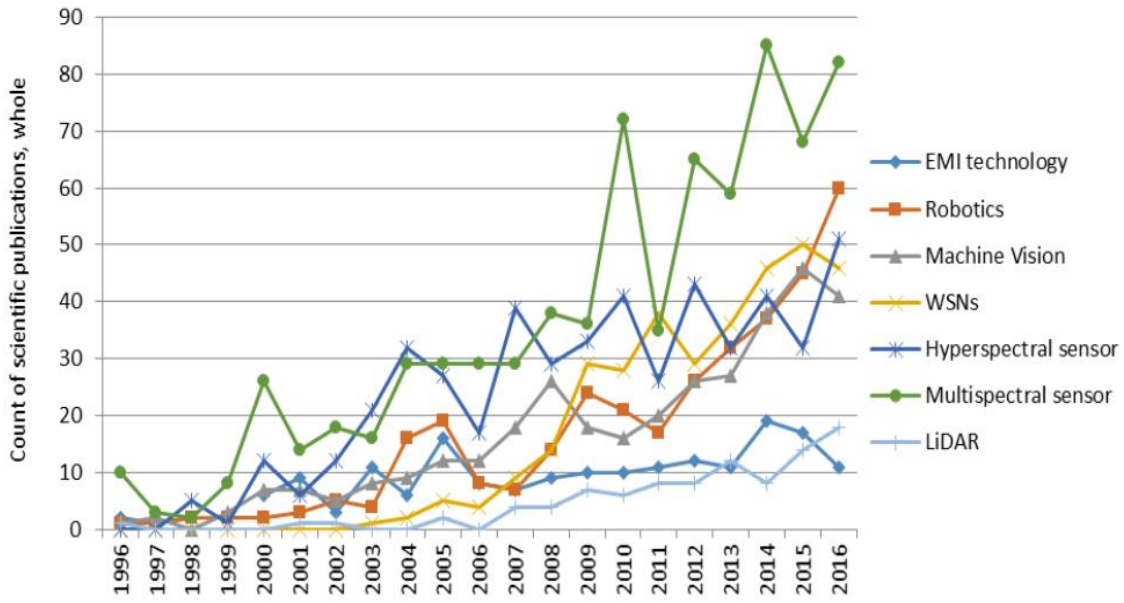


Figure 1-1 Increase of publications in precision farming technologies between 1996–2016 (Cushnahan et al., 2017) such as Electromagnetic Induction (EMI) and Wireless Sensor Networks (WSN). Some technologies such as LIDAR are used in GIS and research via digital terrain models and contribute to Precision Agriculture research without specifically having papers written on their use.

1.4 Background concepts

1.4.1 Precision farming

Precision agriculture or precision farming is a farm management concept around optimising the use of inputs and improving outputs. The term Precision Agriculture was first introduced in 1990 as the title of a workshop (held in Great Falls, Montana, the US, sponsored by Montana State University) by Pierre C. Robert (1941-2003), a professor at the University of Minnesota's Department of Soil Science. Robert and others actively promoted the idea of "site-specific crop management (SSCM)" or "farming by soil" (Lal & Stewart, 2015, p. 4) and suggested that variable management practices should be applied according to sub-field soil variability. In 1996, the workshop was officially named the International Conference on Precision Agriculture (ICPA). Precision agriculture and site-specific crop management have been used interchangeably in many research papers. However, precision agriculture has a broader meaning than SSCM, covering a broader set of agronomic practices such as soil sampling procedures, fertiliser and

chemical rate selection, the timing of applications and machinery selection (Robert, 1993). McBratney et al. (2005) noted that the focus of precision agriculture would evolve to the management of product quality and the environment as technology advances, rather than simply "farming by soil".

Whelan (2011, p.3) noted, "Precision agriculture, in its current SSCM form, has a long history of innovators and pioneers with a single aim of improving agricultural management". He defined precision agriculture as "a philosophy aimed at increasing long term, site-specific and whole-farm production efficiency, productivity and profitability while minimising unintended impacts on the environment", and SSCM as "a form of precision agriculture whereby decisions on resource application and agronomic practices are improved to better match soil and crop requirements as they vary in the field". The US House of Representatives in 1997 defined precision agriculture as "integrated information- and production-based farming system that is designed to increase long term, site-specific and whole-farm production efficiency, productivity and profitability while minimising unintentional impacts on wildlife and the environment" (Zarco-Tejada et al., 2014).

McBratney et al. (1997, p. 141) pointed out "application of the theories of precision agriculture to the practicalities of broad-acre farming relies on the successful handling of the ramifications of uncertainty in the information." This highlights precision agriculture as a data-driven approach for informing decision making in agricultural production systems. Cook and Bramley (1998) explained that by controlling inputs precisely, precision agriculture would increase the likelihood of benefit (such as reducing fertiliser use and improve yield) and decrease the risk of detrimental effects (such as reducing nitrate and pesticide leaching). They illustrated precision agriculture management in a 4-stage circle (Figure 1-2):

- A farmer may observe some low yielding areas on yield maps.
- By measuring exchangeable sodium percentage in the soil, this issue can be interpreted as poor water availability in the root zone.
- Then a potential solution by applying gypsum is evaluated,
- The circle continues after implementing this solution by gathering new information.

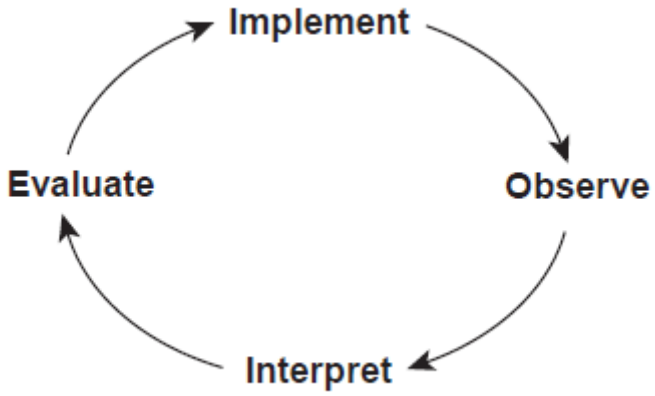


Figure 1-2 A 4-stage cycle of increased precision (Cook & Bramley, 1998)

To narrow the scope of this research, precision farming is strictly defined as a philosophy to manage subfield variability by fine-tuning the crop management inputs such as seed, lime and fertilisers. Because this research is funded by FAR, it focuses on the precision farming application on arable production.

1.4.2 Geospatial yield monitoring

Geospatial yield monitoring is primarily used in grain production, but also forage and bulk crops such as beets and potatoes. In a grain production system, geospatial yield monitoring measures crop yield during the harvest using a system fitted to the combine harvester. The system combines different sensors to measure the harvested grain mass flow, moisture content, and speed to determine total grain harvested. Yield is then derived from these parameters. Yield monitoring works alongside a differential global positioning system (DGPS) to record yield and other spatially variable information across a field such as geographic location (latitude and longitude), time and elevation of each yield data point. This allows a grain yield map to be created, which provides information on spatial yield variability and supports management decisions for producers (Lal & Stewart, 2015).

Geospatial yield monitoring is an essential component in precision farming and is often considered as the first step in the uptake of precision farming practices (Lowenberg-DeBoer & Erickson, 2019). This is because of two important aspects:

1. Yield monitors collect spatially dense data with a relatively low cost, allowing characterisation of spatial and temporal yield variability. Based on the yield variability, variable rates of crop management inputs such as seed and fertilisers may be applied to match the yield potential at different locations of the field, using specialised equipment.

2. Yield monitor data and yield maps provide a tool for evaluating the effectiveness of the management such as calculating and contrasting spatial gross margin before and after undertaking management. Yield maps can also be used to guide field scouting, design soil sampling schemes, and calculate nutrient requirements to inform variable-rate fertiliser applications.

However, the use of yield data and yield maps in commercial crop production remains a challenge. For example, post-processing of data is often required to remediate various systemic and human errors (reviewed in section 2.3.1.1.1) to correlate yield data with other data layers. Also, given the temporal variations of yield, a single-year yield map on its own has limited value for management due to its inability to expose the long-term underlying causes of yield variability associated with soil and seasonal weather patterns. Yield maps collected from multiple years tend to show a more stable yield pattern, providing a basis for spatially varying yield goals for crop management within-field. FAR has been working with maize growers in the North Island of New Zealand to retrieve the yield monitor data from their combine harvesters. Some of this data has been provided by FAR to be analysed as part of the research presented in this thesis.

1.4.3 Site-specific crop management zones

One key aspect of precision farming or SSCM involves delineating management zones (MZs), which are defined as "sub-regions of a field that express a homogeneous combination of yield-limiting factors for

which a single crop input is appropriate" (Doerge, 1999, p. 2). As compared to the traditional whole-field management approach, in which each field is treated as a homogeneous area, SSCM considers the spatial variability in soil and topography within each field and customises inputs according to that variability. This SSCM approach should result in a more efficient application of inputs, which can reduce agronomic, economic costs and hidden environmental costs from potential unused nutrient losses to the environment through leaching, runoff, and gaseous emissions (Bongiovanni & Lowenberg-DeBoer, 2004).

The delineation of MZs has been largely constrained by the available data collection methods that are often time-consuming and expensive to undertake. For example, for fertiliser application recommendation, the spatial variability of soil fertility has been recognised through georeferenced soil sampling and soil test results, which suggests a need to conduct more intensive spatial soil sampling and vary application rates to improve fertiliser application efficiency (Kaul & Grafton, 2017). In New Zealand, soil tests are expensive and cost around NZ$50 per standard test for soil NPK (Ravensdown Ltd, 2019). New Zealand arable farmers often spend NZ$10/ha per year on soil tests before planting crops and most cannot afford high-resolution soil sampling for detailed nutrient mapping. This situation is being improved by advances in sensor technology, making large data sets available for delineating MZs and interpreting yield variability (Kamilaris et al., 2017).

MZs can be delineated based on soil properties such as soil electrical conductivity (EC) and soil organic matter (OM) content, as well as crop sensing and remote sensing imagery. Other common approaches included yield mapping followed by elevation difference across a field (Khosla et al., 2010). Farmer knowledge and past management experience are also proven as a valuable input in delineating MZs (Fleming et al., 2004; Martínez-Casasnovas & Arnó, 2018). Zonal sampling can be undertaken for each MZ that informs variable rate applications of inputs such as seed and fertilisers, for these zones. For example, significant differences in soil fertility may result in the potential to vary the rates of fertiliser between different MZs. The inputs can then be applied automatically across the field using a variable rate

technology such as a precision planter, leading to a potential improvement to input-use efficiencies and cost-savings (Figure 1-3).

These current advances in technology are gradually turning MZ maps into commercially viable agricultural products for large-scale adoption. Based on a MZ map that is produced digitally, the specific rate of crop input can be applied to each MZ using computer-controlled variable-rate technologies (VRT), leading to SSCM (Figure 1-3). For example, maize precision planters can alter the seeding rate on the go automatically based on prescription maps developed in GIS mapping software to determine the population to be planted in different MZs in the field (Holmes & Jiang, 2018).

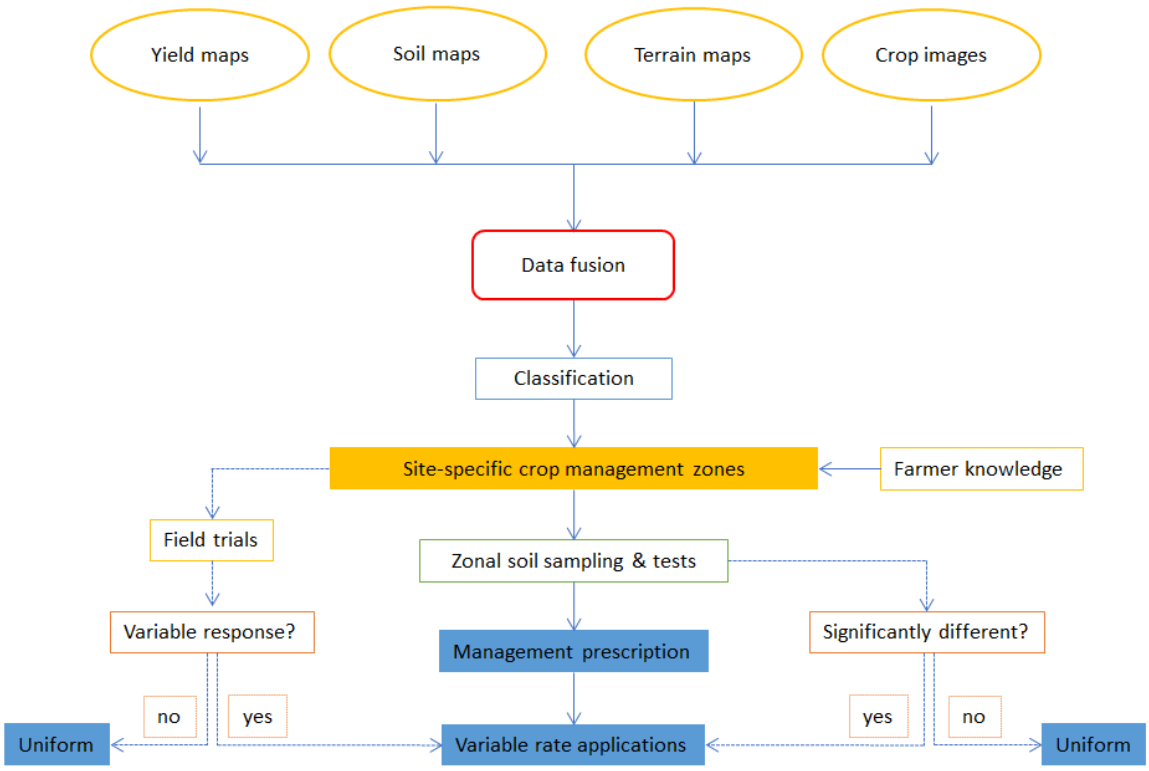


Figure 1-3 Conceptual framework of SSCM (solid lines represent the flow of information in SSCM; dotted lines represent the validation processes required to justify the effectiveness of the MZs delineated)

1.5 Principles adopted

Spatial information derived from field sensors needs to be handled and systematically processed with the help of computers and Geographic Information Systems (GIS). GIS has now become an essential tool for handling spatial data in many aspects of agriculture, from farm management and resource conservation to a broad range of agribusiness applications. Advances in computing power make it possible to analyse and utilise large, high-resolution data sets.

With the increase of computer power over the years, the advent of machine learning algorithms has provided new opportunities to extract useful information from sensor-derived data. Machine learning is a subfield of artificial intelligence, defined as giving a machine the ability to learn from the data without being explicitly programmed. There are two major types of machine learning: supervised and unsupervised. Supervised learning describes a technique that involves a model extracting knowledge by learning from single or multi-variate relationships based on single or multiple factors in the dataset and making predictions using labelled data. Unsupervised learning describes a technique that involves a model extracting knowledge by identifying similar patterns in single or multivariate space based on unlabelled data.

The delineation of MZs is generally achieved using unsupervised learning techniques such as clustering, density estimation and self-organising map, which all partition the data into different sub-regions (Guastaferro et al., 2010; Leroux et al., 2018; Pantazi et al., 2015). The MZs delineated are often separated by some hard and static boundaries (therefore called "static MZs") based on within-field soil variabilities such as soil water and nutrient content.

However, this "static" approach emphasises long-term within-field variability but is unable to address the temporal variability of yield from one year to another (Fraisse et al., 2001). In countries with more variable climates such as New Zealand, farmers may not experience a consistent yield pattern from one year to another. This highlights the need to emphasize the temporal and dynamic nature of yield within-field in New Zealand arable systems.

Therefore, this research seeks to predict spatial yield by incorporating meteorological data such as precipitation, solar radiation and air/soil temperature, which are two fundamentals for crop growth, into the machine learning model. Once the functional relationship between crop yield potential (the response variable Y) and these attributes (the predictors X_1, X_2, X_3, …) is established, the models should be able to make yield predictions at the subfield level (Figure 1-4). There has been considerable interest in applying supervised machine learning techniques for accurate crop yield prediction and for estimating nitrogen requirements for crop growth (Chlingaryan et al., 2018). However, the number of studies focusing on predicting yield variability at the subfield level has been very limited.

FAR has developed an online data management system called ProductionWise® (https://productionwise.co.nz/) from 2018 to help farmers keep records and FAR advisors to communicate with farmers. However, ProductionWise® mainly focuses on keeping farm and field records with no predictive power. The data used in this research and the prediction models could be additionally integrated into this system to provide spatial yield predictions and encourage variable rate applications to be undertaken using precision farming technologies.

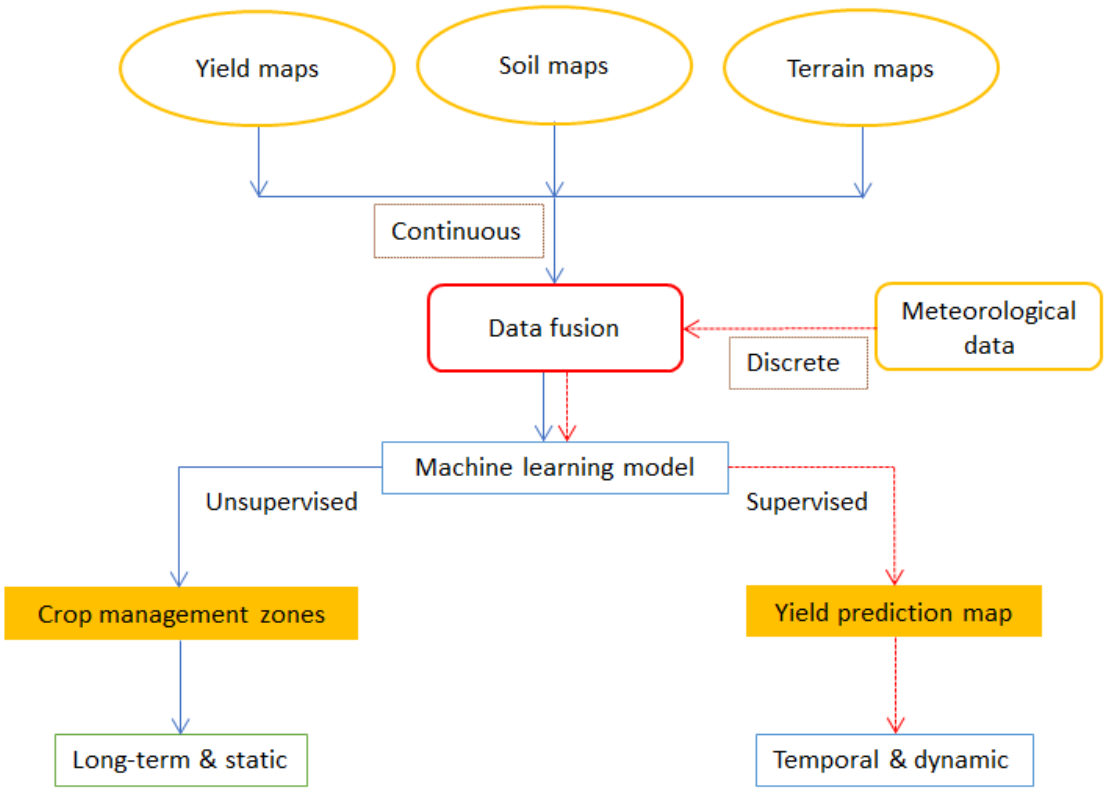


Figure 1-4 Proposed analysis framework by machine learning (blue lines represent the conventional use of the data for delineating MZs; the red lines represent the proposed analysis method)

1.6 Research aim

Yield is the main driver of revenues for arable farmers. Establishing zones where revenues and expenses can be managed and optimised according to field yield potential or site-specific crop management zones (MZs) is a key concept in precision farming. However, little has been done to find appropriate methods to delineate zones for practical use in New Zealand from a variety of spatiotemporal information for crop management.

This research aims to help address this issue by providing a practical and scalable method to identify MZs using machine learning techniques and existing sensor-derived data, to determine if the spatial yield can be predicted for delineating dynamic management zones.

This research hypothesized that:

> The process of delineating site-specific management zones can be improved by modelling spatiotemporal interactions between spatial crop yield and other complementary factors.

To achieve this aim, this research is divided into several objectives:

- To develop a filtering algorithm to improve maize yield mapping precision.
- To identify appropriate spatiotemporal yield predictors by examining historical yield maps, delineating subfield management zones, and undertaking soil sampling.
- To determine the viability of predicting dynamic maize yield at the subfield spatial scale using supervised machine learning algorithms.

Meeting these aims will provide insights into managing field variability and the ability to improve the efficiency of on-farm resource use. Major inputs, such as seeds and fertilisers could be better utilised, and long-term production sustainability and profitability could be improved.

1.7 Thesis overview and chapter outline

This thesis includes the following chapters:

Chapter 1 introduces the aims and objectives of this research. This research is funded by FAR and centres on how to combine field sensor technologies and artificial intelligence for predicting yield variability within-field, which may benefit crop management prescription such as mid-season fertiliser application.

This research seeks to answer this main question:

> Is it possible to improve the process of delineating site-specific management zones by modelling spatiotemporal interactions between spatial crop yield and other complementary factors?

Chapter 2 reviews previous studies and identifies the issues associated with yield sensor data collection and remediation, applications of sensor data and validation, and statistical methods for mapping spatial point data and developing site-specific crop management zones.

Chapter 3 describes the study location, data collection and analysis methods used for this research. Several study sites with consistent management histories were selected. To investigate the yield variability within-field, yield monitor data over several years was collected, filtered, mapped and delineated into zones. To explain soil variability, soil core sampling was undertaken based on a delineated zone map, followed by soil texture analysis. Satellite multispectral images (RGB and NIR) were collected and used to verify the effect of soil variability on crop and yield. Temporal data (rainfall, solar radiation, soil temperature) collected from nearby weather stations were incorporated into the process of predicting spatial yield. Then several machine learning algorithms (SMLR, FFNN, CART, RF, XGBoost and Cubist) were implemented to predict yield and evaluated in terms of prediction accuracies.

Chapter 4 describes the results of the analyses undertaken in Chapter 3. The quality of yield maps after spatial filtering was evaluated using geostatistics. The delineated MZ maps based on yield, soil EC/elevation and crop images showed similarity in their spatial patterns and strong associations between this information. The results of soil texture analysis reveal potentially significant spatial variation in

subsurface features across a short spatial distance and were found highly positively correlated to the soil EC at the soil surface. The modelled results demonstrated the viability of predicting spatial yield at the subfield scale based on inputs from a mix of spatiotemporal variables, which could lead to delineating dynamic MZs and more precise prescription of crop management inputs.

Chapter 5 discusses the results presented in Chapter 4 and explains what they mean for precision agriculture use in arable crops in New Zealand.

Finally, Chapter 6 presents the main findings of the study and the limitations of the work and makes suggestions for future research.

1.8 Work published

Holmes, A., & Jiang, G. (2017). Precision agriculture for New Zealand potatoes–effect of variable yield, tuber size and income. *In Proceedings of the 7th Australasian Conference on Precision Agriculture, Hamilton, New Zealand.*

Holmes, A., & Jiang, G. (2017). Effect of variable rate lime applications on autumn-sown barley performance. *Agronomy New Zealand*, *47*, 37-45.

Holmes, A., & Jiang, G. (2018). Increasing profitability & sustainability of maize using site-specific crop management in New Zealand. In *Proceedings of the 14th International Conference on Precision Agriculture, Montreal, Quebec, Canada*.

Jiang, G., Grafton, M., Pearson, D., Bretherton, M., & Holmes, A. (2019). Integration of precision farming data and spatial statistical modelling to interpret field-scale maize grain yield variability in New Zealand. In *Proceedings of the GeoComputation 2019, Queenstown, New Zealand*.

Jiang, G., Grafton, M., Pearson, D., Bretherton, M., & Holmes, A. (2019). Integration of precision farming data and spatial statistical modelling to interpret field-scale maize productivity. *Agriculture*, *9* (11), 237.

Jiang, G., Grafton, M., Pearson, D., Bretherton, M., & Holmes, A. (2021). Predicting spatiotemporal yield variability to aid arable precision agriculture in New Zealand: a case study of maize-grain crop production in the Waikato region. *New Zealand Journal of Crop and Horticultural Science*, 1-22.

Chapter 2 Literature review

2.1 Introduction

Increasing crop farming expenses, environmental concerns and the uncertainty of commodity markets have driven arable farmers to search for greater input-use efficiency as well as alternative crops, markets, and production systems. One of the ways of improving input use efficiency is to vary the application of inputs such as irrigated water and fertilisers according to soil or crop requirements. This is known as precision agriculture or precision farming. Crop farmers have informally practised the philosophy of precision agriculture for thousands of years. Farms have traditionally been divided into individual field parcels based on soil and topography, in which each field is treated as an homogeneous area (Oliver, 2010). However, because GPS was not available until the 1980s, crop management was practised at the individual field scale. Sub-field variability was often left unmanaged, which became more apparent when farmers merged their fields into larger units to accommodate centre pivot irrigation or larger farm machinery (Blackmore et al., 2005). The practices of site-specific crop management (SSCM) are necessary to improve input-use efficiency such as irrigation and fertiliser use to minimise environmental risks (Basso et al., 2016).

This chapter examines problems associated with the use of geospatial information and reviews previous studies that use sensor-derived information for crop management (section 2.2).

The focus and scope of this review include:

- Issues associated with yield sensor data collection and remediation (section 2.3.1.1.1 and section 2.3.1.1.2);
- Applications of sensor data and validation (section 2.3.1);
- Statistical methods for mapping spatial point data and developing site-specific crop management zones (section 2.3.2).

2.2. Using spatial data to prescribe management inputs

Modern-day precision farming employs data-driven approaches for informing agricultural production system decisions. One key aspect of precision farming involves delineating potential site-specific crop management zones (MZs). The emphasis is on separating "sub-regions of a field that express a homogeneous combination of yield-limiting factors for which a single crop input is appropriate" (Doerge, 1999, p.2). The outcomes of this management method include more efficient use of inputs, which can reduce agronomic, economic, and hidden environmental costs, from potential unused nutrient losses to the environment (Bongiovanni & Lowenberg-DeBoer, 2004).

The following sections (2.2.1 and 2.2.2) review precision farming applications based on MZs and their limitations and then identify the research problem in the analysis of spatial data.

2.2.1 Optimising seeding rates

The cost of seed is one of the major expenses of growing maize in New Zealand (Genetic Technologies Limited, 2019). Variable-rate seeding (VRS) has the potential to improve profit by matching the seeding rate to production potential within the field. Variable-rate seeding is a technique used to adjust seeding rates spatially within a field based on soil texture and/or other variables as the soil in some parts of the field will have different soil moisture-holding capacity that supports differing seeding rates (Jeschke et al., 2015). Once the soil variations within a field are characterised, the information can be processed to determine suitable seed rates and then uploaded to a variable rate planter, which adjusts the planting density automatically. This process is not new to maize growers, but there are limited studies on the value of VRS in the public domain.

Shanahan et al. (2000) investigated the yield response of different maize hybrids (early and late maturity) and plant densities (24,692, 37,037, 49,382 and 61,727 plants/ha) in dryland landscapes of the US Great Plains during 1997, -98, and -99 seasons. The treatment yields and landscape data (such as elevation and slope) were interpolated and then classified into low-, medium-, and high-yielding areas based on the average and standard deviation for each field. This study found large within-field yield variability in the

three fields (average yields ranging from 5.43 to 6.39 t/ha) and the coefficient of variation (CV) ranging from 20% to 29%. Hybrids responded similarly to field variation while plant densities responded differentially in the low, medium and high-yielding areas (R^2 = 0.43, 0.76, 0.87, respectively). Economically optimum plant densities changed by around 5000 plants/ha between high and low-yielding field areas, producing potential savings in seed costs of US$6.25/ha. This study suggested that there is a potential cost-saving from the use of VRS based on different landscape attributes (such as elevation and slope) that influence the spatial yield.

Taylor et al. (2006) evaluated a VRS trial of wheat on a commercial farm in New South Wales, Australia. The field was divided into MZs based on soil electrical conductivity (EC surveyed using a Geonics EM38 and a Veris 3100 sensor) and RTK elevation. To verify MZs, soil sampling was conducted with core samples taken from two depths (0-30 cm and 60-90 cm) at 15 locations within each zone. Nine small trial plots (100 m × 30 m) including three seed treatments (50, 75 and 125 kg/ha) and three replicates were randomly located in each zone, with the remaining areas planted with 100 kg/ha as a reference. The zones did not appear to have significant variations in the measured soil properties, except for the subsoil CEC and the percentage of silt in the topsoil ($p<0.05$). In different zones, variable seed rates produced different responses in the plant counts and the harvest yield, which suggested the potential for VRS. However, only one year's seeding trial was conducted, which did not have enough temporal resolution to make conclusive results. Therefore, VRS based on spatial management zones may have to consider temporal factors such as forecast rainfall and soil temperature, in addition to soil texture and crop cultivar to ensure optimal seeding rates.

Hörbe et al. (2013) attempted to optimise the maize plant population according to management zones in two fields (90 and 124 ha) in Southern Brazil. For Experiment 1 (the 2009/10 season), the field was delineated into low, medium, and high-performance MZs visually based on the farmer's knowledge. For Experiment 2 (the 2010/11 season), the MZs were delineated based on multiple yield maps. A random block design was used for the experiment, with five seeding rates (50, 60, 70, 80, and 90 thousand

seeds/ha) and five replications set up in each zone. Based on the two experiments, reducing the recommended seeding rate by 31% in the low MZ resulted in a yield increase of approximately 1.5 t/ha. Increasing the recommended seeding rate by 13% in the HZ resulted in an increase of 0.91 t/ha in grain yield. However, the experiments did not consider the effect of temporal conditions such as rainfall on the yield potential, so is not conclusive as only one growing season was investigated.

Licht et al. (2017) developed procedures for maize seeding rate optimisation and maximizing yield using soil and topographic parameters. Experimental treatments included five seeding rates (61,750; 74,100; 86,450; 98,800; and 111,150 seeds/ha) in a randomised complete block design in three central Iowa fields from 2012 to 2014 (nine site-years). Soil samples were analysed for available phosphorus, exchangeable potassium, pH, soil organic matter (SOM), cation exchange capacity (CEC), and texture. Topographic data (in-field elevation, slope, aspect, and curvature) were determined from publicly available light detection and ranging (LIDAR) data (Dubayah & Drake, 2000). Multiple regressions with interaction terms were conducted to understand and identify the key independent variables that best explained maize grain yield. However, this study found no consistent interaction between the measured variables and the seeding rate for most sites due to weather variability. These results further confirm that the performance of variable rate seeding is likely to have an interaction with weather conditions during the growing season.

There appears to be a need to develop a crop management strategy that can incorporate weather variables into the decision making processes about prescribing within-field seeding rates.

2.2.2 Optimising nitrogen (N) fertiliser input

Nitrogen (N) is the most frequently applied nutrient to arable crops as it has a rapid effect on crop growth. N is often the limiting factor to crop growth as it is readily leached through the soil profile in the nitrate form. Also, N can be volatilised as ammonia, nitrogen gas, and nitrous oxide. More N than is needed by the crop is often applied, as the cost of additional N is low compared to the potential yield loss resulting from insufficient N. As farmers have traditionally not been accountable for the consequences of nitrogen leaching or greenhouse gas emissions, there has been little incentive to manage its use responsibly. There

are now nitrogen limits and seasonal use restrictions imposed by regional Government bodies, highlighting the importance of managing on-farm nitrogen use, and providing the potential to apply variable rates of nitrogen within-field (Basso et al., 2016).

Godwin et al. (2003) examined revenues from precision farming practices during a three-year study in Southern England with known farm sizes and levels of variability. VRA-N in barley and wheat were evaluated based on historic yield and shoot density. Historic yield trials were conducted over three years in three fields using both variable and uniform application rate strategies. Shoot density (the number of shoots per square meter) was determined in near real-time using NDVI data from airborne digital photography. The use of VRA-N based on real-time crop canopy sensing was also evaluated in winter wheat. An average benefit of £29.90/ha (NZ$55/ha, 2020 conversion rates) was reported from using the shoot density method, and £22/ha (NZ$42/ha) from using the crop canopy method, compared to the standard, uniform N application rate. However, this analysis is inconclusive because it overlooks many temporal factors such as rainfall and does not have enough data layers (three years) to investigate temporal yield variation due to VRA-N.

Koch et al. (2004) evaluated the economic feasibility of four VRA-N strategies utilising MZs with an on-farm field study in Colorado, US. Trials were conducted over three site-years on two long-term maize fields in north-eastern Colorado under furrow (18.5 ha) and centre-pivot (58 ha) irrigation during the 2000 and 2001 growing seasons. The management zones for the two fields were delineated based on bare soil aerial imagery and farmer knowledge of topography and experience.

The N management strategies were:

- Uniform with a constant yield goal;
- Grid sampling-based to meet a constant yield goal;
- Site-specific management zone – constant yield goal (SSMZ-CYG);
- Site-specific management zone – variable yield goal (SSMZ-VYG).

The N rates were calculated using an N recommendation algorithm, with the inputs being expected yields, soil nitrate, and soil organic matter. This study found consistently less total N fertilizer (6% – 46%) was used with the SSMZ-VYG strategy when compared with uniform N application. Net returns from the SSMZ-VYG strategy were US$18.21 (NZ$28, 2020 conversion rates) to US$29.57/ha (NZ$45) more than uniform management. The cost savings of SSMZ-VYG were consistently higher than the grid sampling-based strategy. In this study, despite that the variable yield goal was subjectively defined (based on the farmer's knowledge and past management experience rather than using yield monitor data), the resulting cost-saving suggested that there is a value for setting the variable yield goal.

Basso et al. (2013) evaluated the impact of variable rate nitrogen (VRA-N) fertiliser application on spatial and temporal patterns of yield in a durum wheat grain crop. The study was conducted during the 2008/09 and 2009/10 growing seasons in a 12-ha field near Foggia, Italy. The field was subdivided into two MZs being High and Average, using yield maps from the two seasons. Three N rates were identified using a calibrated crop simulation model (SALUS). They were: low N (30 kg N/ha); average N (70 kg N/ha); and high N (90 kg N/ha), This study found no significant effects of the different N rates for the 2008/09 growing season for the High and Average zone. For the 2009/10 growing season which experienced higher rainfall there was a significant difference between the three N rates in grain yield for the Average zone (2,955 kg/ha), but not in the High zone (3,970 kg/ha). This study suggests the optimal amount of N for a given MZ varies with the rainfall amount and distribution during both the fallow and growing season, which makes the prescription of N for a rain-fed crop more difficult. Therefore, the delineation of MZs based on spatial information such as historical yield maps should ideally combine temporal factors, such as rainfall, to prescribe the input of N more effectively.

2.3 Addressing the issues

Section 2.2.1 and section 2.2.2 identified issues regarding the lack of temporal focus for delineating MZs using sensor-derived spatial data. Although the process of delineating static MZs is generally well-

established and representative of yield or soil variation, the spatial pattern of the yield could be dynamic and variable over time. It is difficult to prescribe optimum rates of input without knowing how much yield each area will produce in the current season.

To fill this knowledge gap, this research aims to improve the delineation of MZs by predicting spatial yield within-field using data that is high spatial resolution and readily available from the fields. Once the relationship between yield and these spatiotemporal factors is established, it may be possible to prescribe crop management inputs at the sub-field scale.

2.3.1 Sensor-derived data

To prescribe variable-rates of fertiliser within-field based on soil fertility, the cost of undertaking intensive soil sampling using traditional sampling methods such as grid soil sampling is expensive. The advent of precision farming sensors such as soil electrical conductivity (EC) sensors show great promise for collecting good resolution spatial data with minimal costs of time and labour. The use of sensor-derived spatial data could provide insights into predicting spatiotemporal patterns of crop yield.

The common spatial data available to New Zealand arable farmers include:

- Yield monitor data derived from combine harvesting operation (section 2.3.1.1);
- Soil electrical conductivity (EC) derived from mobile soil EC sensors (section 2.3.1.2);
- Elevation data derived from precision planting operation (section 2.3.1.3);
- Multispectral imagery derived from satellites (section 2.3.1.4).

2.3.1.1 Yield monitor data

Site-specific yield data from crops are recorded using a yield monitor on a harvester. This is achieved by several subsystems: product output sensing; area sensing (ground speed and swath width); positioning (DGPS); data processing, monitoring, and storing units. There are two types of yield monitor systems on the market: grain volume (volume-flow sensors) and grain mass (mass-flow sensors). Yield monitor

systems all provide a final output of crop yield in tonnes per hectare (t/ha) at a user-set time interval, usually 1 second (1 Hz) while harvesting (Demmel, 2013). The area covered is calculated by multiplying the working width of the cutter bar and the ground distance travelled by the harvester between each period.

2.3.1.1.1 Yield data errors

It is necessary to apply post-processing to all yield monitor data before mapping variability to eliminate any spatial artefacts caused by systemic or human error. To remediate these errors in a GIS software program and to reduce misleading information, this section reviews possible errors that may exist in some yield data.

Systemic error

1) Delay of the grain flow signal which records the mass or volume at a different location to where harvested: When grain flows through the combine it experiences conveying, threshing, separation and cleaning processes before finally reaching the yield monitor. A "simple time delay model" was introduced to match the flow data to the position data, and the flow signal is shifted by a constant time value (Birrell et al., 1996).

2) Combine filling mode: Many yield maps exhibit low yields immediately inside the headlands where the combine first enters the crop, which is caused by combine filling, which is due to the sieves, elevators and other temporary storage spaces being empty and taking time to fill up, before passing through the combined mechanism and reaching the yield sensor) (D. Murphy et al., 1995; Thylén & D. Murphy, 1996). In that case, the low yield recorded could be caused by a shortcoming of the monitoring system, rather than any actual variation in the crop.

Human error

1) Incorrect set width: When the combine harvester enters the crop, the operator must align the edge of the head cutter to the crop rows, the cut width is generally set as the full width of the header. However, a problem may occur at the end of harvesting a field when there are only a few rows left and these are less than the full cut width of the machine. An unknown crop width

entering the header and or the actual crop width entered incorrectly into the controller. This causes long strips of low yield to be recorded running along the length of the field, where the harvester finished working by only harvesting a fraction of the full head width (Blackmore & Marshall, 1996).

2) Inconsistent ground travel velocity can be caused by operator behaviour such as difficulties in aligning the crop divider with the crop edge, and abrupt changes of combine ground speed (Colvin & Arslan, 2000). For example, a sudden decrease in ground speed would result in a small estimated area at a given time, making the instantaneous yield too large as the grain keeps travelling in the combine.

3) Empty header: Another issue is that while the combine is turning around in the headland, there are times when the header is lowered, and recording yield, but is not harvesting (i.e. no grain flow through the clean grain elevator). This creates invalid data points (Luck & Fulton, 2015).

Sensor calibration

Measurement errors can be mitigated by statistical filtering, but actual yield information is lost if the yield sensor is not adequately calibrated. By sensor calibration, the recorded yields are adjusted based on the actual yield measured on a certified weight scale. Ideally, at least four grain-loads should be measured to establish a reliable calibration curve (Luck & Fulton, 2014). However, during a busy harvest schedule, the calibration of the yield sensor is often poorly carried out using fewer loads. Instead of the “four-point” as recommended by the sensor manufacturer, a linear “two-point” calibration method may have been undertaken. That is, farmers would measure one load at the high grain flow scenario and then measure another load at the low grain flow scenario, simulated through the change of ground travel velocity or cut width (Luck & Fulton, 2014). If the yield data is calibrated from the “two-point” method, it is likely to have recorded higher values than the actual yield as high levels of measurement error tend to occur in the low grain flow condition (Kormann et al., 1998). Therefore, it is suggested for the best practice use of yield

data to follow the recommended sensor calibration method and sensor maintenance at the beginning of harvest to minimise the errors.

2.3.1.1.2 Spatial data filtering

Given potential errors in yield measurement as reviewed (section 2.3.1.1.1), undertaking data filtering of spatial data is important before its use in decision making.

There are a few studies on how to detect yield data errors and remove them in GIS programs. An algorithm developed by Simbahan et al. (2004) includes two stages of yield data screening: The first stage addresses flow delay and start-/end- pass; and the second stage undertakes outlier tests (velocity, grain flow, and moisture), yield thresholds, local neighbour filtering and overlay points. Inverse distance weighted (IDW) interpolation (Burgess & Webster, 1980) was used to estimate yield within a search window and this was then compared with the actual yield. If the actual yield was outside the confidence interval of the estimated yield, the value was classified as an outlier. For the validation of this algorithm, 1000 points of maize and soybean yield were withheld as test data and compared to that interpolated by ordinary kriging with root mean square errors (RMSEs) calculated. The results showed that the RMSE was reduced by 4.34 – 5.36 % after the screening, which indicates the accuracy of the yield maps increases with each screening.

A freeware package, Yield Editor, constructed by Sudduth and Drummond (2007), implements several filters to remove flow delay, start- and end- pass delay, velocity, swath width, yield and positional errors and allows users to visualise the effect on an interactive map display. Yield Editor 2.0 incorporates an ``automated yield cleaning expert'' module for the automated selection of value thresholds based on the median range of the values of all the points contained in the dataset (Sudduth et al., 2012). Their study reported 13% to 27% (2510 to 8325 points) removal from the maize and soybean yield data collected from six fields. However, this program is not integrated into mapping software and is limited to the data from specific yield monitoring systems.

Spekken et al. (2013) proposed a generic approach to filter spatial data, which simply requires two input parameters: a search radius less than the range of spatial dependency; and a specified maximum coefficient of variation. Without the requirement for any further information (e.g. time, ground velocity), it is more effective in reducing noise (fill mode and lag time, points near the headlands, and points with an erroneous set width, which represent the majority of errors) on an interpolated yield map than filtering the data based on yield thresholds. They reported that 29% of the maize yield raw data points were removed by the programme.

2.3.1.2 Soil electrical conductivity (EC) data

2.3.1.2.1 Correlations between soil EC and soil texture

Electrical conductivity (EC) is a measure of the ability of the soil to conduct electricity, commonly expressed in units of milli-Siemens per meter (mS/m). Electrical current may be conducted through the soil via three pathways: 1) the pore-connected soil solution of water and ions; 2) the cations that are bound to the surfaces of clay particles; 3) connected solid soil particles (Corwin & Lesch, 2003; Rhoades et al., 1999). The soil EC measurement can be correlated with soil properties that affect crop productivity, such as soil texture and cation exchange capacity (CEC).

Sudduth et al. (2003) evaluated EC measurements obtained from a non-contact, electromagnetic induction (EMI)–based sensor (Geonics EM38) to those from a coulter-based sensor (Veris 3100) on four fields in Illinois and Missouri with fine-textured soil. The EC data were used to relate to soil particle size fractions and CEC sampled from 12 to 21 sampling sites in each field to a 120 cm depth, within a single field and measurement date. The strongest correlations were found between EC and clay content ($r = 0.6–0.9$) and CEC ($r = 0.51–0.88$), which were consistent across all four fields and sensor types. Data obtained with both types of EC sensors were similar and exhibited similar relationships to soil physical and chemical properties. Similar results were also derived from their later study (Sudduth et al., 2005) in the north-

central US, which reported the highest correlation (R^2 = 0.37-0.63) between EC and topsoil clay content across 12 fields.

Moral et al. (2010) produced prediction maps of soil texture from a handful of soil samples (N=70) using regression kriging with soil EC (measured by a Veris 3100 sensor) as secondary information for a 33-ha rapeseed field in Badajoz, south-western Spain. The results found that the soil EC was positively correlated with clay (r = 0.67 – 0.77), and negatively correlated with coarse sand (r = -0.61 – 0.67) and fine sand (r = -0.63 – 0.69).

In Australia, Rodrigues Jr et al. (2015) predicted soil clay content and CEC using an EMI survey (soil EC) and gamma radiometric soil survey data (Dierke & Werban, 2013) in six fields ranged from 25 - 90 ha across three sugarcane growing regions. Their models derived from principal component analysis (PCA) were able to predict CEC (adjusted R^2 = 0.29-0.71) and clay (adjusted R^2 = 0.42-0.73) at most sites. These findings suggest it is possible to use soil EC data to map soil variability, which is considerably less expensive than when undertaken by traditional sampling methods such as grid sampling.

In New Zealand, Hedley and Yule (2009) used high-resolution EC maps (derived from Geonics EM38 survey) to predict the spatial variation of soil water content in a 33-ha irrigated maize field in Palmerston North. Three zones with low, intermediate, and high EC values were delineated based on the EC map. Higher EC values were associated with higher silt (30.8 to 61.3%) and clay content (8.3 to 26.8%) for the top 60 cm soil, respectively. Within each zone, EC values were correlated (R^2 = 0.80) with a range of volumetric soil water contents to develop a relationship between EC, soil texture, soil moisture, and different soil moisture deficits. The soil EC map, with real soil moisture monitoring, can then be used to inform variable-rate irrigation within-field to prevent nutrient leaching due to over-application (Hedley et al., 2010). Hedley's papers demonstrated that EC can be used as a predictor for soil texture and soil water holding capacity and soil CEC. However, the emphasis of their research was the management of irrigation application on large fields (> 20 ha) based on soil variability. There is insufficient emphasis on the effect of spatial soil variability on crop yield in predominantly non-irrigated arable fields (<10 ha) in New Zealand.

2.3.1.2.2 Correlations between soil EC and crop yield

Lund et al. (2000) used a ``boundary line" approach to derive yield goals with soil EC maps and multiple season yield data. The analysis divided the EC values into several bins and then fitted with a quadratic model (R^2 = 0.96). However, only the static pattern of yield potential was derived and the temporal relationship between soil EC and yield potential from one year to another is not well understood.

Kitchen et al. (2003) modelled within-field yield potential based on the soil EC measured by a Veris 3100 and topographic data (slope, curvature, and aspect) in three contrasting soil–crop systems (four site-years of maize, three site-years of soybean, and one site-year each of grain sorghum and winter wheat) in Colorado, Kansas, and Missouri during 1997–99. Using either regression or neural networks analysis, EC alone explained yield variability (averaged over sites and years R^2 = 0.21) better than topographic variables (averaged over sites and years R^2 = 0.17). Combining EC and topography measures improved the model R^2 values (averaged over sites and years R^2 = 0.32) for the non-irrigated fields in Kansas and Missouri. However, the low R^2 produced by the models suggested the need to include other field attributes such as soil fertility, climate and management factors (such as the amount of fertiliser) for estimating yield potential.

Kitchen et al. (2005) investigated the use of soil EC and elevation for delineating MZs for two claypan soil fields planted in a maize-soybean rotation in central Missouri. Meanwhile, the historical yield maps (between 1993 and 2002) were grouped into ``deficit", ``optimal" or ``excessive" precipitation years and were delineated into the respective yield MZs for these temporal conditions. The MZs were then compared with the yield MZs derived. 40–60% agreements (κ = 0.21 - 0.43) were found between different yield zones and soil MZs. The highest agreements (58-60%; κ = 0.29 – 0.34) were found with the yield zones derived from yield maps of the "deficit" years, whereas relatively lower agreements (41%-46%) were found with the yield zones derived from yield maps of the "optimal" years. These results suggested that the effect of soil texture on yield is more pronounced in a relatively drier year.

Guastaferro et al. (2010) delineated MZs (relatively high yielding and low yielding) using kriged soil texture data (to 0.3 m depth) and potential yield over three seasons for a 12-ha durum wheat field in south-eastern Italy. Overall, weak areal alignment (20%-40%) was found between MZs and yield classes. The results also indicated that the MZs did not have consistent associations with yield classes due to temporal variations, consistent to Kitchen et al. (2003; 2005).

The studies in this section suggest that despite crop MZs being able to identify different yield levels, they are unable to provide potential yield values of each zone, making it practically difficult to prescribe variable input rates. While some studies seek to associate yield with soil EC using a robust boundary line modelling, machine learning techniques such as neural networks and geostatistical technique such as regression kriging, little research has been done with sufficient focus on modelling the spatiotemporal interactions between yield, weather data (such as rainfall) and EC. The lack of temporal focus in these studies raises the question of how would yield respond to management inputs such as N-fertiliser under different temporal moisture conditions. To prescribe N rates more effectively based on within-field soil variability, it may be feasible to obtain a prediction of the yield response using historical data, which would allow management such as mid-season N fertiliser applications to be undertaken.

2.3.1.3 Topographic features

Soil at a higher elevation within a field may have shallower topsoil and comprises a coarser texture than soil at a lower elevation, which can have an impact on drainage (Rumbal, 1978). These factors can control the way water accumulates or moves across, and within, the soil (Whelan & Taylor, 2013, p. 129). The aspect of a sloping soil surface can have a direct effect on yield variability. Depending on the direction of the predominant weather systems in a region, there may be differences in growing conditions due to aspect. Geary (2003), using the CERES wheat model found a loss in grain yield of 1 t/ha on a slope of 10 % oriented to the North (away from the sun) in England. In the Southern hemisphere, North- and West-facing slopes are warmer and dry more quickly than those facing South and East. In New Zealand, arable

farming is predominantly undertaken on small blocks and relatively flat land, the effects of slope and aspect on yield are often less pronounced. However, it is increasingly common to produce arable crops on sloping fields as they are planted using minimum tillage technologies to reduce the risk of soil erosion and nutrient runoff (Haynes & Knight, 1989). In these circumstances, the effect of elevation on spatial yield may exacerbate soil moisture and fertility distribution. RTK-GPS is by far the most reliable method to record elevation, with a precision of 12 to 20 cm.

2.3.1.4 Spectral data for crop monitoring

Plants use the energy from sunlight to produce glucose from carbon dioxide and water, by photosynthesis. Plants have little ability to absorb near-infrared (NIR) light in the 700 - 1300 nm wavelength from sunlight radiation. Light absorption by chlorophyll is particularly higher in the blue (400 to 500 nm) and the red (600 to 700nm) part of the visible light spectrum than in the green region (500 to 600nm) (Whelan & Taylor, 2013, p.58). This results in plants having a higher reflectance of green light. In healthy growing crops, a large disparity exists between low red light and a high NIR light reflectance. Red reflectance increases when crops are under stress because the chlorophyll reduces activity (Huete et al., 2004). However, NIR reflectance decreases as the leaf internal structure collapses; and thus, the disparity between the reflectance decreases significantly (Adams et al., 1999; Gausman et al., 1976; Tucker, 1979).

Reflectance measurements of the green, red, and NIR spectrum has been used to determine plant nitrogen (N) content and canopy N deficits (Buschmann & Nagel, 1993). Techniques were developed for using a SPAD chlorophyll meter, colour photography, or canopy reflectance factors, to assess spatial variation in N concentrations in crops.

Blackmer and Schepers (1995) suggested the use of images of canopy reflectance centred at 550 nm acquired late in the growing season could be used to detect portions of the field that were nitrogen deficient. In the US Great Plains, an index was developed from green and NIR reflectance of an irrigated maize crop, which was highly correlated with an N sufficiency index calculated from SPAD chlorophyll

meter data and provided a rapid assessment of maize plant N status (Bausch & Duke, 1996; Bausch & Diker, 2001). There are several indices calculated from the light reflectance, of which one of the earliest and most robust measures is Normalised Difference Vegetation Index (NDVI) (Jackson & Huete, 1991; Wiegand et al., 1991). Higher values of NDVI indicate greater vigour, health or greenness of the crop. Strong correlations (R^2 = 0.7 - 0.92) were also found between different vegetation indices (NDVI, transformed soil-adjusted vegetation index (TSAVI), green NDVI) acquired at mid-grain filling and maize-grain yield during a two-year trial in US Nebraska (Shanahan et al., 2001).

However, even if it is possible to predict maize-grain yield using remote sensing data collected during mid-grain filling, it is likely too late to undertake any management intervention at that point (Pinter Jr et al., 2003). For example, for a maize crop, nitrogen fertiliser is normally applied at planting and around mid-November when the crop is at the V6 (6th leaf vegetative) stage, i.e. 4-5 weeks after planting. Mitigating the impact of nutrient stress on crop yield through the application of N fertiliser ought to occur at earlier growth stages. Failure to apply N in this optimal application window may result in irreversible loss of potential yield. This means that even if enough fertiliser is given at a late stage such as at grain fill, it is unlikely to reverse the yield loss resulting from insufficient N at an earlier stage. Also, if variable-rate fertiliser application is desired, then the current application methods make it difficult to apply inputs at a later stage of the season, because the crop may be too tall for a tractor to pass through without causing significant damage to the plants.

2.3.1.4.1 Reflectance-based management zones

De Benedetto et al. (2013) combined multisensory data including soil EC (Geonics EM28DD sensor) (surveyed in May & September 2010); vegetation indices (NDVI, NIR/Green ratio, Red Edge) obtained from Worldview-2; and GeoEye-1 satellite imagery for MZ delineation on an irrigated tomato field in southern Italy. The field was split into two blocks for separating two irrigation treatments from July 15, 2010: optimal water supply and moisture deficit conditions. The study concluded that the clustered maps based

on multi-sensor information can be used to differentiate the effect of crop moisture stress under different irrigation treatments.

Martínez-Casasnovas and Arnó (2018) compared different potential MZs (NDVI-based, NDVI-EC-based, NDVI-EC-elevation-based) delineated based on the accumulated-NDVI derived from Sentinel-2 multispectral images, soil EC derived from Veris 3100 sensor and elevation data for a 45-ha field of sunflower-maize rotation under centre pivot irrigation in north-eastern Spain. An alternative to MZs was created by combining the farmers' expert knowledge about the within-field soil characteristics such as soil texture and problems such as drainage with the accumulated-NDVI map. The farmers' knowledge accurately reflected statistically different levels of yield when compared with partially successful methods of delineating MZs based on NDVI, EC and elevation. However, only single-year yield data from one irrigated field was used for validation.

Georgi et al. (2018) developed an automatic segmentation algorithm for developing MZs based on multi-spectral satellite data (RapidEye time-series images from 2009 to 2015) for wheat and canola fields in north-eastern Germany. The algorithm automatically selected suitable images for cloud identification and crop pattern delineation based on a set of statistical thresholds for the blue, NIR band reflectance and NDVI. The NIR bands of all selected images were averaged and divided into five classes (quantile 10%, 35%, 65%, 90%). The result showed that the delineated MZs were consistent with the expected level of yield productivity (relative yield <85%, 85%-97%, 97-103%, 103%-107% and >107%) or yield expectancy zones (very low, low, average, high and very high) related to soil texture variations within-field. This study suggested that there is potential for using reflectance images captured by satellites to delineate management zones when historical yield maps are absent, which may verify the effect of soil texture variations on crop growth.

Ekanayake et al. (2018) used reflectance data (RGB pixels) derived from historical Google Earth images to develop management classes for a 5.95 ha maize field in Pukekohe, NZ. MZs derived from the bare-soil-only images better depicted the patterns in the traditional multi-year yield map (R^2 = 0.6 – 0.99),

compared to the map derived from the crop-cover-only images (R^2 = 0.62 - 0.96). However, further ground validations should be undertaken to verify the cause of yield variability and determine if there is a correlation with the crop reflectance.

The biggest challenge for applying remote sensing techniques to arable farming in New Zealand is the acquisition of consistently good quality satellite images that can be related to the appropriate stages of crop growth. To monitor the maize crop in New Zealand, the probability of obtaining cloud-free images is high mainly from December to February (summer), but low before this period (spring). This makes it difficult to estimate yield when cloud-free images are not available for the relevant specific growing stages (Eberhardt et al., 2016).

Currently, high-resolution remote sensing products (< 10 m) are priced based on the area covered (often costing between NZ$20 -NZ$60 per km^2 with a specified minimum order area) (FAO, 2015). This means that their cost is not justified financially when being used for management at the field level. Research needs to demonstrate the value of images for precision farming, which will encourage the adoption of zonal soil sampling and mapping, delineating management zones.

2.3.2 Statistical techniques for analysing spatial data

2.3.2.1 Geostatistical interpolation and Kriging

Field sampling of soil and crop properties with GPS (e.g. Yield monitor, soil EC) collect data as spatial points, which need to be interpolated into surface maps to be utilised by VRA technologies such as a precision planter. That means predictions of the soil property in question could be made at regular intervals across the area of interest, represented as a raster grid. When the grid is of a moderate to high resolution (i.e. when the distance between the centres of adjacent grid cells is small, maybe 100 m or less) digital soil mapping techniques can provide detailed spatial coverage of the soil variable of interest (McBratney et al., 2003). In the absence of covariates or predictor variables, the predicted value of the

target property at any location is solely dependent on the spatial configuration of the set of locations where the target property value is known.

Geostatistics is a technique for interpolating spatial point data based on Tobler's First Law of Geography: ``Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 234). This spatial relationship is quantified mathematically using a variogram (or semivariances), which illustrates that the difference in values increases as the distance increases before eventually reaching a plateau (meaning the difference in values is no longer associated with the distance). The distance in which the semivariance increases is called the "range of spatial dependency" or "range", which can be used to guide the sampling interval (Kerry & Oliver, 2003).

Kriging is an advanced interpolation technique that is widely used to produce maps based on underlying spatial variations between sampled locations (J. Li & Heap, 2014). As with IDW (Burgess & Webster, 1980), the kriging estimate of the target soil or crop property at an unobserved location is a weighted average of the values of the property at the surrounding, observed, locations. The difference is in how the weights are determined. In the inverse distance weighting (IDW), the weights applied to the computation at any unobserved site are based solely on the distance of the unobserved site to the observed sites $D_E(x_0, x_i)$. In kriging, the weights are based on not only the distance to the observed sites, but also the modelled variance in the target property between them (J. Li & Heap, 2014). When data are enough to compute variograms, kriging is often used as an interpolator for sparse data.

2.3.2.2 Spatial classification for delineating management zones

The delineation of MZs for VRA is an important aspect of precision farming practices. The statistical techniques for undertaking MZ delineation have evolved along with the fields of artificial intelligence and machine learning (ML). During the last 15 years, there have been researching interest in machine learning (ML) techniques for accurate crop yield prediction and nitrogen status estimation using rapidly advancing precision farming data (Chlingaryan et al., 2018). By establishing functional relationships $f(X_1, X_2, X_3 \dots)$

between spatial crop yield potential (the response variable Y) and related attributes (the predictors X_1, X_2, X_3 …), a machine learning model can be built using the ``split-sample'' approach (supervised learning) to make predictions of spatial yield variability or to identify spatial patterns without any prior knowledge of the data (unsupervised learning). For delineating MZs within a field, unsupervised learning techniques such as fuzzy c-means clustering are mainly based on single or multiple factors that explore sub-regions with similar yield-limiting factors such as soil water and soil nutrients (Guastaferro et al., 2010; Leroux et al., 2018; Pantazi et al., 2015).

2.3.2.2.1. Unsupervised classification

Fuzzy c-means is an unsupervised classification algorithm (unsupervised classification is a term in machine learning used to describe identifying patterns without any prior knowledge), which aims to minimise the objective function through numerous iterations.

Fuzzy c-means clustering has been embedded in several software packages such as Management Zone Analyst and FuzME (Fridgen et al., 2004; Minasny & McBratney, 2002) and is used extensively to partition data points into groups for delineating MZs (Lark & Stafford, 1997; Kitchen et al., 2005; Taylor et al., 2007; Moral et al., 2010).

Lark and Stafford (1997) introduced fuzzy c-means clustering to interpret multiple-year yield maps and identify patterns of seasonal yield variation. The fuzzy clustering produces partial membership, a measure of statistical confidence in which one data point may be associated with multiple clusters (rather than belonging to only one cluster) (Bezdek, 1973). This technique was tested on yield data (1993-95) of a 6-ha, uniformly managed winter barley field in England. Soil samples were taken from 100 locations at 0-20 and 20-80 cm depths across the field and measured for moisture, OM, pH and mineral nitrogen. An Analysis of Variance (ANOVA) test was used to determine the randomness of soil properties within the clusters. The ANOVA results suggested that the between-season yield variation in a cluster may reflect water deficiency in some soil types (Lark & Stafford, 1997).

Fraisse et al. (2001) aimed to delineate MZs based on soil EC and topographic attributes such as elevation and slope using fuzzy c-means clustering. Data collected in two claypan fields (28-ha and 36-ha) located in central Missouri were used to test the proposed methodology. The determined optimum number of MZs, indicated by the sum of within-cluster yield variances, were inconsistent from year to year and were possibly a function of seasonal weather and the crop species planted. The number of zones decreased if adequate moisture conditions were present throughout the cropping season, or if crop species tolerant of water stress were planted.

A protocol was developed for delineating potential management classes based on multiple-year yield data for broad-acre crop production in Australia (Taylor et al., 2007). The protocol used mostly free software such as Yield Editor (spatial filtering), VESPER (variogram and kriging) and FuzME (fuzzy c-means clustering). The field average kriging variance was used to calculate 95% confidence intervals to determine the optimal number of clusters (Whelan & McBratney, 2003). The results suggested that some soil physical properties such as soil depth and plant available water capacities were significantly different between different yield classes delineated using the protocol.

Like Fraisse et al. (2001)'s approach, Blasch and Taylor (2018) developed a Multi-temporal Yield Pattern Analysis (MYPA) method. Principal Component Analysis (PCA) was performed to produce the principal component (PC) maps that represented static and dynamic patterns of the yield. The outlier maps from the raster stack were identified using PCA. Yield pattern stability was evaluated using statistical per-pixel analysis based on the standard deviation (SD) of normalised yield. MZs were delineated with important PCs (PCA loading > 0.5) and SD using k-means clustering. However, this analysis is only applicable to complete yield datasets from a whole field due to the limitation of PCA, which is not capable of handling missing data due to management factors such as changing field boundary or trialling different treatments or hybrids in the same field.

These papers provide different alternatives to analyse multiple-year yield maps and delineate potential productivity zones. However, these static maps identified long-term yield variation but did not examine

the dynamic pattern of the yield within each zone. These maps are perhaps suitable to be used in some continental countries where stable weather conditions are expected from year to year. In New Zealand, the weather conditions could be more variable between months, between weeks or even between days, than in the continental countries, which limits the effectiveness of static zones and subsequently variable rate applications. Therefore, temporal conditions needed to be integrated into the process of delineating management zones to estimate yield potential or provide decision support as to how much input should be applied for each zone within a field at a time.

2.3.2.2.2. Supervised classification

Statistical models (such as stepwise multiple linear regression and partial least square) have been used to help understand the relationship between crop yield and measured soil and site parameters, using large, spatial, multivariate datasets. Correlation and other linear techniques (e.g. multiple linear regression models) have been used in many previous studies to explore the relationship between crop or soil properties and the data derived from precision farming tools (Blasch et al., 2015; Schirrmann et al., 2011; Stadler et al., 2015). The correlations can also provide insight into the linkages between precision farming data and crop yield spatial variability (Drummond et al., 2003; Kitchen et al., 2003; Wang et al., 2019). However, linear regression models assume that the relationships between the dependent and independent variables are linear. Improved results have been reported with more complex machine learning techniques such as artificial neural networks.

Blackmore et al. (2003) identified the spatial and temporal trends using crop yield maps from four different fields over six years in England. On assuming mostly stable yielding pattern within-field over the years due to the same yield-limiting factors, they attempted to predict the spatial yield of the latest year using yield maps from previous years but found poor results (R^2 = 0.02 – 0.43). Their results could suggest that a multivariate approach should be undertaken for predicting spatial yield.

J. Liu et al. (2001) used a feed-forward, back-propagation neural network model for predicting maize yield based on soil factors (e.g. soil texture, pH, phosphorus, potassium, organic matter), management factors (nitrogen fertiliser), and monthly rainfall as inputs in their neural network. This study was conducted on small plots with different fertiliser treatments in the Morrow Plots of the University of Illinois. The BPNN was able to model the interaction between rainfall and the rate of applied nitrogen fertiliser. It also predicted maize yields with 80% accuracy. However, only one algorithm was tested. Given on-going development of ML algorithms in recent years, better predictions may be achieved with more advance ML algorithms such as deep neural networks.

Drummond et al. (2003) predicted maize and soybean yield on three sites in the US state of Missouri (ranging from 13 to 36 ha in size) also using a feed-forward, back-propagation neural network with several soil fertility parameters (e.g. soil pH, OM, phosphorus, calcium, magnesium, potassium) and topographic inputs (e.g. elevation, slope). The neural network generally provided better statistical predictions (an average of 45% of the variation, range 21 - 74%) in the multiple-year analysis than the other models (SMLR, which explained an average of 31% of the yield variation, range 14% to 65%). However, variable-rate application based on intensive sampling was also not cost-effective to be used for managing sub-field scale variation due to high sampling and mapping costs.

Therefore, this research attempts to provide a case study on predicting spatial yield within-field using different machine learning models based on a variety of spatiotemporal information on maize fields in New Zealand, which is major cropping system in the North Island. It is hypothesized that this case study can provide new insights to help to manage within-field crop and soil variability by prescribing variable rates of inputs such as N fertilisers within-field. The application of sensor-derived data such as yield monitor data and soil electrical conductivity (EC) offers advantages over techniques that use spatial data collected from intensive and expensive grid sampling. The minimal costs associated with sensor-derived data are more likely to be of commercial interest to New Zealand crop farmers and therefore will encourage further uptake of variable rate application.

2.4 Discussion and Conclusions

This chapter identifies the issue in the analysis of spatial data for managing within-field yield variability using VRA. Even though farmers have collected yield data over many years (in some cases up to 25+ years), the uptake of precision farming practices in New Zealand, such as calibrating yield sensors, yield mapping and delineating management zones, is limited because of a lack of a temporal focus on prescribing crop management inputs within-field.

Specifically, this review has revealed the following research gaps:

- Yield monitor data collected with GPS can be mapped to present the spatial variability of crop yields within-field. However, there are potential issues with the sensing systems such as combine flow delay and human variability such as inconsistent ground velocity, which may be detrimental to how well the yield maps represent the actual yield. Several spatial filtering techniques have been developed, including Yield Editor which is widely used to filter out yield data errors. However, no standard method was established for spatial filtering and some of the programmes were not straightforward for data that is collected using different sensors that come with different formats. To support yield mapping in New Zealand, spatial filtering should be flexible with different sensors and should be integrated into the rest of the yield data analysis to help improve the accuracy of yield maps.

- To derive within-field yield potential, the soil EC data measured by EMI (such as EM38) or direct contact sensor (such as Veris 3100), along with RTK elevation have shown some promise. High EC values are often associated with fine-textured soil and greater available water holding capacity, which generally prevents crop moisture stress. The implication is to alter crop management inputs spatially, for example by lowering input rates in areas with low yield potential, which may result in lower production costs or produce a better yield. Of particular interest is that the impacts of soil texture variations on spatial yield may be captured using high-resolution satellite imagery such as Worldview-2, GeoEye-1 satellite imagery and RapidEye, which can be used to delineate

management zones for a range of crops when yield maps are absent. Strong correlations to crop yield have also been reported with several variants of NDVI such as Soil Adjusted Vegetation Index (SAVI) and Green NDVI. However, studies on the use of publicly free, medium-resolution satellite imagery such as Sentinel-2 for subfield-level crop research have been scarce.

- Previous studies have attempted to delineate MZs using historical yield maps or soil EC and/or topography. However, spatial patterns delineated from yield or soil-based MZs are static, whereas the yield responses within the MZs can vary temporally. Most MZ studies do not consider temporal factors such as meteorological data and are unable to estimate yield potential for each MZ. Without a detailed level of understanding of yield potential and crop response to specific variables (e.g. climate, soil type), it is difficult to apply variable rates of input to MZs or justify its financial performance. Therefore, this research and its stated aims exemplify a need to predict within-field yield potential for informing variable rate applications such as; N fertiliser by exploring spatiotemporal interactions between crop, rainfall and soil temperature using appropriate machine learning (ML) techniques.

Chapter 3 Materials and methods

3.1 Introduction

This chapter describes the study area and sites, data collection, and analysis methods employed for the research so that the objectives stated in Chapter 1 can be achieved:

- To address Objective 1 (improve maize yield mapping precision), a customised spatial filtering algorithm was developed and tested using yield monitor data over four years (2014, 2015, 2017, and 2018) collected from FAR's NCRS (section 3.3.1).
- To address Objective 2 (identify appropriate spatiotemporal yield predictors), soil electrical conductivity (EC), soil organic matter (OM), RTK elevation, and multispectral crop images were collected from FAR's NCRS and correlated with yield (section 3.3.2, section 3.3.3). Zonal soil sampling and soil texture analysis were undertaken to attempt to explain the yield variability and to provide a calibration for the use of soil EC (section 3.3.5).
- To address Objective 3 (to determine the viability of predicting spatial yield at the subfield spatial scale), five non-irrigated maize fields with consistent management history in the Waikato region were selected (section 3.3.6). The temporal yield variability, meteorological data (rainfall, solar radiation and GDD) were incorporated for predicting subfield yield variability, which can induce 'dynamic' production yield zones from year to year (section 3.3.4). Several machine learning models (SMLR, FFNN, CART, RF, XGBoost, and Cubist) were implemented and evaluated in terms of their yield prediction accuracies (section 3.4.5).

These models can potentially be embedded into GIS software and then used by farm consultants to help inform crop management decisions such as N-fertiliser application.

3.2 Study area and site description

3.2.1 Study area

The study area is in the Waikato region of the upper North Island, NZ, and is home for a population of 472,100 (Statistics NZ, 2018a). It has a wide range of agricultural industries, agribusinesses, and research institutes. The region is best known for its grass-fed dairy production industry and is a large consumer of grains for supplementary feed when pasture production is below animal demand. Poultry farming is another major industry in the region, supplying both chicken meat and fresh eggs which also creates a domestic demand for grain production (Statistics NZ, 2018b). Crop growing and processing are therefore significant contributors to the region's economy, with 1,485 farms involved in arable cropping in the Waikato (Statistics NZ, 2019). Other crops include herbs, tomatoes, cucumbers, and lettuce. The region has more than 4,000 hectares of land dedicated to growing outdoor fruit and vegetables (Statistics NZ, 2019).

The Waikato region is the largest producer of maize crops (*Zea mays*) in New Zealand (Statistics NZ, 2019), supplying grain or silage primarily as supplementary feed for livestock. Most of the maize hybrids planted in New Zealand are dual-purpose and can be grown for silage or grain. Maize crops are C_4 pathway plants (C_4 plants have different anatomy and physiology compared to C_3 plants and have greater water use efficiency in carbon fixation) and are typically grown in areas of the North Island of NZ due to higher levels of solar radiation and a warmer temperate climate, ensuring a higher yield potential compared to the South Island (Booker, 2009).

The climate of the Waikato region is temperate and tends to be warm and humid in summer and mild in winter (Chappell, 2013). The average annual rainfall is 1,250 mm spread evenly throughout the year (NIWA, 2016). However, 'dry spells' (Periods of fifteen days or longer with less than 1 mm of rain on any day from December to March) are common in the region (Chappell, 2013).

Soils in the area around Hamilton have a high horticulture and cropping potential. Of note is the Horotiu soil, a local Allophanic soil (Typic Udivitrand) formed from a series of thin tephra (volcanic ash) layers

overlying slightly raised channel/bar volcanic alluvial deposits laid down about 18,000 years ago by the Waikato River (Lowe, 2010). A key component of this soil is the amorphous mineral “allophane”, which binds soil particles into stable and fine aggregates that allow free drainage and helps retain moisture. Root exploration is relatively unconstrained by this soil type (Molloy, 1998). The Horotiu soil often forms a complex association with the Te Kowhai silt loam, gley soil (Typic Ochraqualf) formed from “volcanogenic overbank flood deposits” located on the lower-lying alluvial swales formed by the Waikato River.

3.2.2 Site description

To identify appropriate spatiotemporal yield predictors, five long-term (10+ years) non-irrigated maize fields in the Waikato region were selected (Figure 3-1) because of their consistent within-site management history. Yield monitor data were consistently collected from all fields over several years and provided by FAR.

Northern Crop Research Site (NCRS) (175.372 E, -37.835 S) and is located at Tamahere, 10 km south-east of Hamilton, see Figure 3-1. It lies within the alluvial plain landscape unit of the Hamilton Basin where the Horotiu-Te Kowhai soil complex is dominant. An Automatic Weather Station (AWS), administered by New Zealand’s National Institute of Water and Atmospheric Research (NIWA) is located at Hamilton airport, approximately 5 km SW from NCRS, providing hourly weather data for public use. Because of the close distance to the NCRS site, the weather data recorded by the AWS is likely to be an acceptable proxy for the weather conditions at the NCRS site.

The NCRS site also hosts long-term trials for maize and forage, including a long-term crop establishment trial. The selected maize field is a 10-ha in size and has been solely dedicated to growing maize. Strip-tillage has been practised for planting maize at the site since 2013. Strip-tillage is a conservation tillage system for planting row crops and only cultivates a portion of the soil between previous crop rows and keeps 75% of the residue on the soil surface, thereby minimising soil disturbance (Heege, 2013), This contrasts with conventional tillage which normally fully cultivates the soil and mixes crop residue through

the soil before planting. Nitrogen (N) fertiliser (typical urea with a volatilisation inhibitor) is typically applied at planting, and also in the mid-season at the V5 crop stage (five-leaf vegetative stage) via surface broadcasting using a twin-disc spreader.

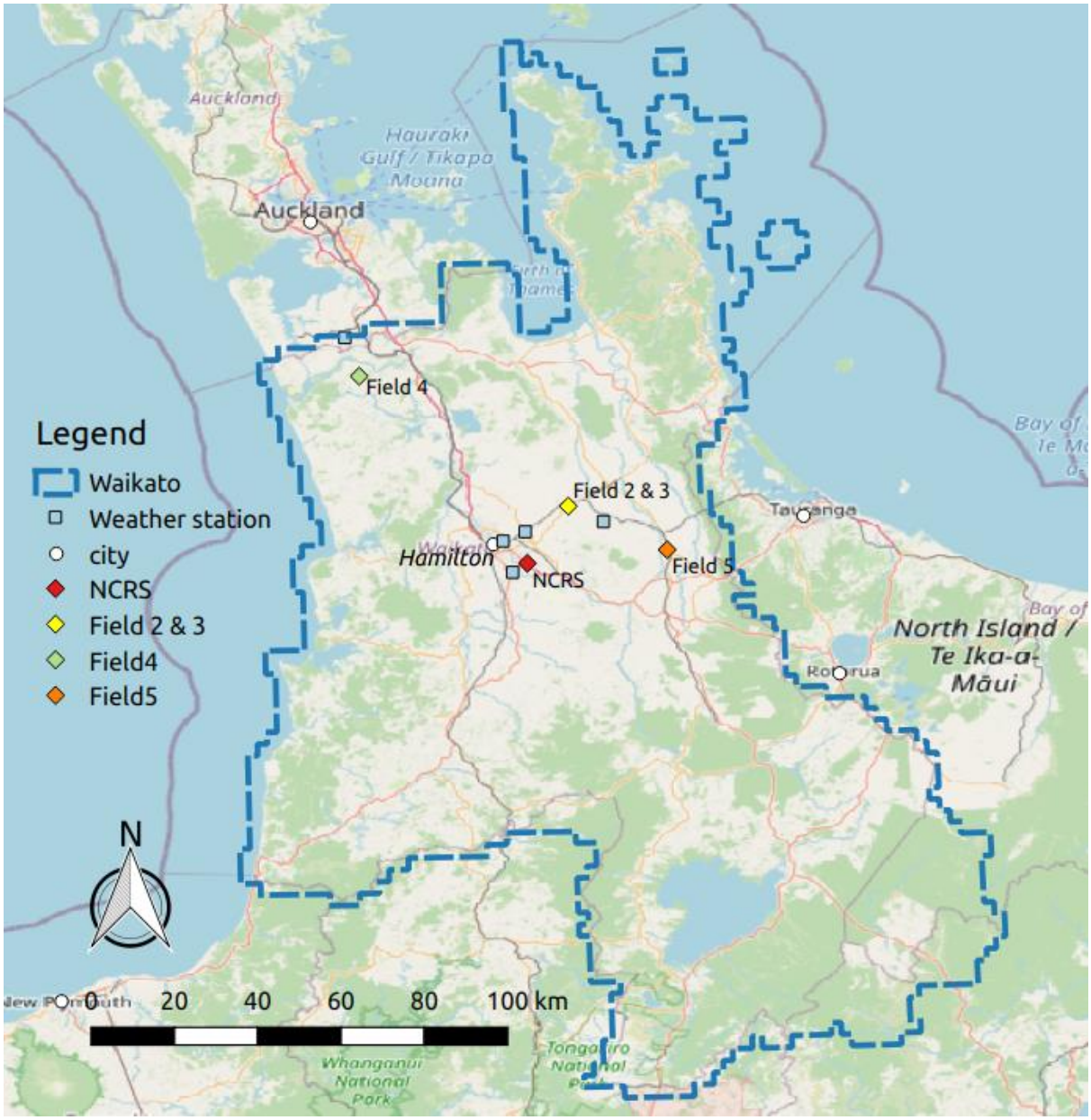


Figure 3-1 Study site location in the Waikato Region of New Zealand and the nearby NIWA weather stations (source: Open Street Map)

3.3 Data collection

3.3.1 Yield monitor data (response variable)

Spatial yield data of maize (*Zea mays*) grain was collected from four harvests (2014, 2015, 2017, and 2018) at NCRS during crop harvesting using a yield monitor (Figure 3-2) fitted on an 8-row (6 m swath) combine harvester with a GPS receiver. Spatial data points were recorded at 1-second intervals during harvest. As an example, the first ten rows of 2018 yield data subset are shown in Appendix 1.

The principle of geospatial yield monitoring was described in section 2.3.1.1. The yield monitor records mass grain flow during a series of combine harvesting processes (cutting, mixing and threshing). An on-board computer then converts the flow relative to the area covered by the harvester into yield since the last GPS reference point and records the grain moisture content using an electrical capacitance sensor.

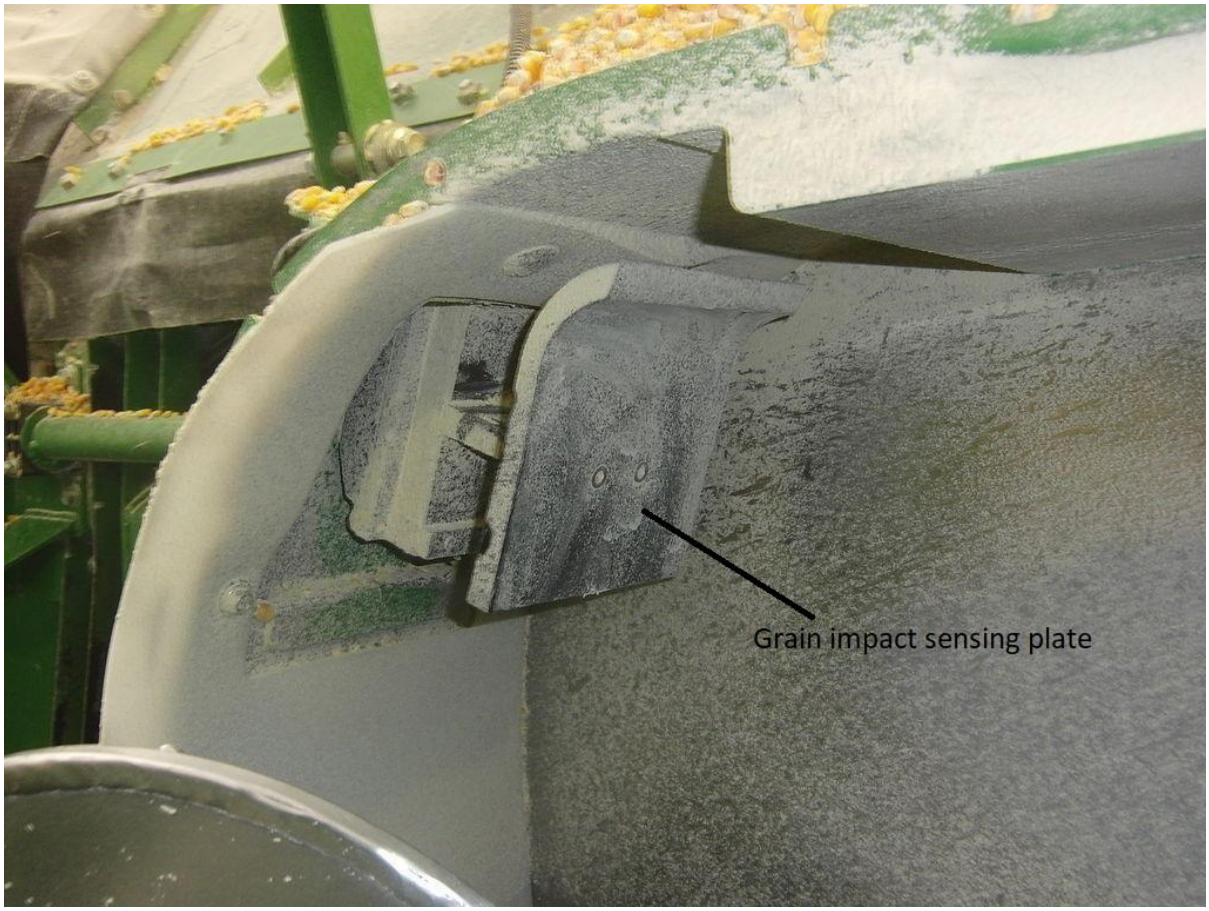


Figure 3-2 Yield monitor based on the measurement of mass grain flow (the sensing plate converts the impact of the incoming grain into electric signals)

3.3.2 Soil EC, OM and elevation (as yield predictors)

Soil apparent electrical conductivity (EC) is related to inherent soil texture and soil water holding capacity which can be used to derive yield potential within-field (Lund et al., 1999). The EC soil data for all available fields were collected using a Veris Mobile Sensor Platform (MSP-3) and collected at 1-second intervals along the transect, with 10 to 15 m separating each transect.

The Veris MSP-3 has three sensing technologies, one of which is a contact-based EC sensor (see Figure 3-3). This system has four-disc coulter plates that measure electrical resistivity indicating how well the electrical current flows in soil. The less resistance to the electrical current through the soil, the higher the moisture content. Differences in EC readings suggest variation in soil texture and/or the presence of a water table. This sensor takes measurements from two soil depths: between 0 and 30 cm (EC shallow), and between 0-90 cm (EC deep), which is predetermined by the voltage and the separation distance between discs (Grisso et al., 2005).

The Veris MSP-3 also has a near-infrared LED that measures the light reflectance from a furrow created within the topsoil. Reflectance values are then calibrated using laboratory testing for soil organic matter content (Hurst et al., 2015; Lund & Maxton, 2011).

Elevation points were measured in October 2017 by an RTK GPS installed on the variable rate planter.

Figure 3-3 Veris Mobile Sensor Platform (MSP-3) (Source: Hurst et al., 2015)

3.3.3 Satellite imagery

To examine if there are any associations between crop images and yield maps, this research explored the use of Sentinel-2 images to map crop variability. Sentinel 2 is a multispectral (443–2190 nm), wide-swath (290 km), fine spatial resolution (10 m for RGB and NIR band) satellite imagery developed by the European Space Agency (ESA) within the framework of the European Union Copernicus programme. Sentinel 2 (Level-1C) has been available since June 2015. The Sentinel-2 Images are free for download and provide reflectance data from the light spectrums (R, G, B, and NIR) with a reasonable spatial resolution (10 m) which is useful for investigating crop performance at the subfield scale. Sentinel 2 has a five-day revisit cycle which is a higher frequency than the older NASA Landsat 8 with 30 m resolution and a 16-day cycle. The shorter revisit cycle of Sentinel 2 increases the chance of obtaining cloud-free images at a crop growth stage for the study sites.

Sentinel 2 images have only been available since June 2015, whilst the yield monitor data for the subject field date from 2014. Therefore, Landsat 8 images are also used for 2014 (Table 3-1).

Table 3-1 Information for the satellite imagery acquired for this research

Satellite Imagery	Acquisition date	Wavebands	Spatial resolution
USGS Landsat 8 Surface Reflectance Tier 1	2014-02-09	Band2 (blue 452 – 512 nm), Band3 (533-590 nm), Band4 (636 – 673 nm), Band5 (851 – 879 nm)	30 m
USGS Landsat 8 Surface Reflectance Tier 1	2015-02-12	Band2 (blue), Band3 (green), Band4 (red), Band5 (NIR)	30 m
Sentinel-2 MSI: Multispectral Instrument, Level-1C	2017-02-23	Band2 (blue 458 - 523 nm), Band3 (green 543 – 578 nm), Band4 (red 650 – 680 nm), Band8 (NIR 785 – 899 nm)	10 m
Sentinel-2 MSI: Multispectral Instrument, Level-1C	2018-02-23	Band2 (blue), Band3 (green), Band4 (red), Band8 (NIR)	10 m

Landsat 8 Surface Reflectance Tier 1 data has been available since April 2013 and has a revisit cycle of 16 days with a spatial resolution of 30 m for the visible and NIR band (Roy et al., 2014). This resolution, however, may be too coarse to be of use at the sub-field scale as some pixels may contain "noise" such as bare soil, streams, and trees within the same pixel (Dash et al., 2017). Remote sensing data (Sentinel 2 and Landsat 8) were pre-processed by selecting cloud-free images and applying atmospheric correction (converts radiance to reflectance). The processed images were downloaded using Google Earth Engine (an online platform that allows records of data from many satellite projects to be accessed and processed). There are several steps included in the selection of images:

1. Cloud-free images were selected based on visual inspection of each image for each acquisition date (Table 3-1). After filtering, only 20 cloud-free Sentinel 2 images were available for the study field between 1 September 2016 and 1 July 2018, with most acceptable images taken between December and August. Georgi et al. (2018) developed an automatic algorithm that can select cloud-free images of vegetation for spatially high-resolution RapidEye satellite multispectral images (5 m). Due to different band and resolution specifications of Sentinel-2 images, this research was unable to reproduce the programme. This aspect of image availability could be explored further in future research.

2. Since this study is focussed on investigating correlations between crop reflectance and soil and yield, images with low vegetation (indicated by low average values of normalised difference vegetation index [NDVI] <0.4) were excluded as they indicate a high incidence of bare soil.

The atmospheric correction of Sentinel 2 images to remove the effect of water vapour, aerosol, and ozone, was undertaken using the "Py6S" program developed by R. Wilson (2013) and incorporated by S. Murphy (2017). The data were exported as a series of raster stacks (a collection of raster layer objects with the same spatial extent and resolution), with each stack containing reflectance data at four wavebands in the R, G, B, and NIR range. To ensure that images with different spatial resolutions could be overlaid with corresponding spatial data, a regular 6 m grid spacing was generated over the raster images to extract pixel values (explained in section 3.4.2).

3.3.4 Meteorological data (as yield predictors)

To determine how yield might respond to changes in temporal conditions at critical stages during growth, meteorological data was downloaded from New Zealand's National Climate Database (CliFlo: NIWA's National Climate Database on the Web), a freely accessible data source. For the NCRS study site, historical weather data was recorded by the Hamilton AWS weather station (175.332 E, -37.861 S), located within 5 km of the site (Figure 3-1).

Weather data was retrieved as CSV files containing daily summary statistics. These variables selected were:

Rainfall: Water is essential for crop development in terms of forming cell structure, germination, photosynthesis, nutrient transfer and transpiration. The water requirement of the maize crop varies between different growth stages but must not be restricted at any stage for maximum yield to be achieved. Theoretically, 20% of total water uptake occurs within the first five weeks of crop establishment, 33% occurs during the next three weeks before silking, and 31% during the next three weeks during silking and early grain fill (Genetic Technologies Limited, 2016). Maize is very sensitive to drought, especially during the two weeks before and after silking, which may cause serious yield losses (McWilliams et al., 1999). Therefore, rainfall occurring during the growing season should be partitioned into different growth stage intervals to model its effect on yield.

Solar radiation: The amount of solar radiation intercepted for photosynthesis during growth determines the crop's rate of growth and potential yield (Muchow et al., 1990). Maize is one of the most efficient plants at converting radiation into biomass (about 1.6 tonnes dry matter/ha for every 100 MJ of radiation intercepted (FAR, 2009). The most sensitive periods of crop growth (such as flowering and early grain fill) are often the most susceptible to stresses such as insufficient light, water, or nutrient intake. Previous studies have found that increases in radiation during grain-filling drives an increase in maize yield, especially in cool-temperate climates such as New Zealand (Yang et al., 2019; D. Wilson et al., 1995).

Air temperature: Temperature drives the duration of growth. The duration of growth is typically evaluated in terms of the accumulated heat unit (metric Growing Degree Day (GDD)), calculated by subtracting the base growth temperature for maize [8°C] from the average air temperature in 24 hours. Maize growth is assumed to be nil at the base temperature (FAR, 2005). In New Zealand, the GDD requirement for maize maturation is about 1,800 GDD and is dependent on the planting date and growing season temperature. The maize maturation date could be delayed for around 30 days if the air temperature was 3 °C cooler for the 12 weeks of spring (Salinger, 1986). A delay of maturity may also create uncertainties for harvest and delays for winter crop rotations.

Daily weather data were aggregated into seven-day intervals, with accumulated rainfall, solar radiation, and GDD calculated over each interval. Because the maize crop was planted on different dates each year, and the duration to reach maturity varies between years, the meteorological data was referenced based on the number of days before and after the date of planting (Day 0) to compare the temporal effects for different years.

3.3.5 Soil core samples

To help explain yield variability within-field, soil sampling was undertaken. The maize field was visited on 19 November 2018. The landscape features were visually inspected and photographed (Figure 3-5). Due to the constraints of time and budget, it was decided to conduct management zone soil sampling (taking samples from each predefined management zone) as compared to grid sampling (taking single or multiple soil cores within each regularly spaced grid cell). The predefined MZs were determined from historical maize yield maps (see section 4.3.2).

Soil was sampled to a depth of 80 cm using a standard soil core sampler (3.5 cm internal diameter) at six locations (Figure 3-4), three each in relatively high yielding zones (HY) and low yielding zones (LY). Each sampling point was georeferenced using an RTK GPS (Leica Zeno 20). Each soil core was then cut into sections (Figure 3-6), sealed in labelled sample bags, and stored in a chiller before processing.

Texture analysis was performed on all samples collected from soil cores at depths of 5-10 cm, 10-15 cm, 15-20 cm, 20-25 cm, and 25-30 cm (Figure 3-6). The soil at 0-5 cm depth was not sampled because of the highly variable bulk density and organic matter content within this interval for most soils (Kaul & Grafton, 2017). This 0 - 5 cm depth is also above the point at which the maize seed was planted (5 cm). Since the crop grows roots downwards in response to gravity, the texture in this 0 - 5 cm depth above planted seed would have minimal effect on the crop. The soil profile below 30 cm was also not measured for soil texture because:

1) Soil moisture deficit (controlled by rainfall, soil texture and groundwater level) before the V6 stage will permanently impair maize yield production after the V6 stage (Song et al., 2019);

2) The rooting depth is about 30 cm deep between the V5 (5th leaf vegetative) to the V6 stage (Abendroth et al., 2011). Previous studies (Nichols et al., 2019) also suggest that 52 to 94% of the total root mass of a mature maize crop was found in the top 30 cm soil, which includes the primary and seminal roots that are critical for nutrient uptake.

All maize seed was planted with coated urea and ammonium sulphate, and additionally, urea was applied mid-season, so hypothetically, all maize plants should have received similar quantities of nutrients. To verify potential variation in nutrient uptake due to variation in soil texture within the field, fifteen-centimetre fixed-transect soil core samples and subsequent testing (for soil fertility status) were undertaken 32 days before planting on 8 October 2018, and again 50 days before the harvest of the crop (14 May 2019). The soil test results were presented in Appendix 4 and 5.

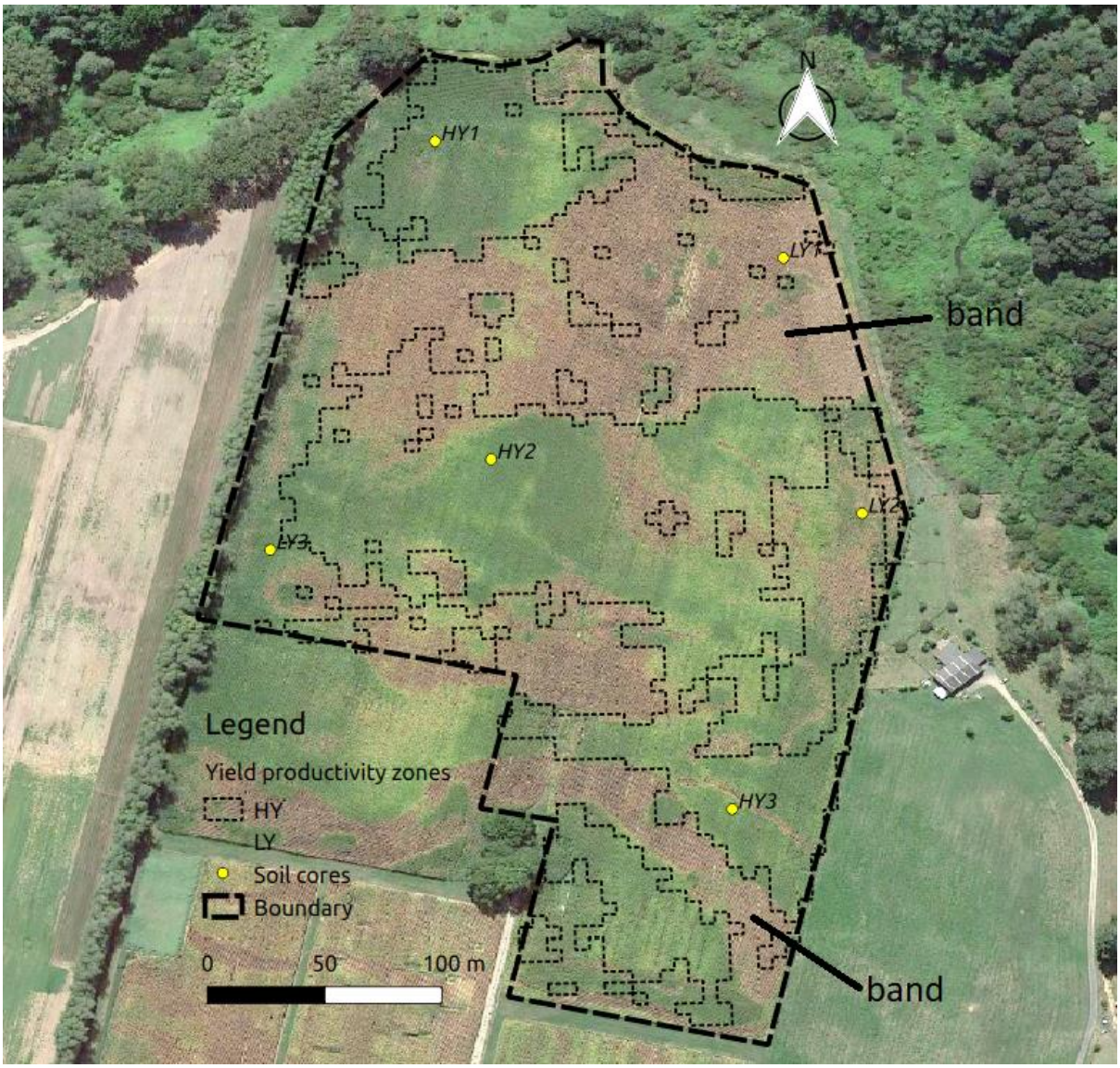


Figure 3-4 NCRS core soil sampling locations and the banding (brown versus green) associated with differential growth patterns (source: Google Earth RGB image; imagery date: 11 March 2016)

Figure 3-5 Photo of the NCRS maize field (looking west) The vertical pole is the location of sample HY3. Note the variation in maize growth in the middle centre and middle right, potentially caused by soil texture differences within the field. (photograph date: 19 November 2018)

Figure 3-6 Core soil sections (0-5 cm, 5-10 cm, 10-15 cm, 15-20 cm, 20-25 cm, 25-30 cm, 30-40 cm, 40-60 cm) taken from the NCRS maize field (photograph date: 19 November 2018)

3.3.6 Other field sites for testing yield modelling

To further test the modelling approach employed in this research (section 3.4.5), four other field sites in the North Island of New Zealand were selected based on the availability of several years of maize harvest data. The sites that were chosen: (a) had a history of long-term maize production, and (b) had consistent management history with intact spatial data.

Site 2 (175.487 E, −37.676 S) and Site 3 are two neighbouring non-irrigated maize fields (5-ha and 14.5-ha, respectively) located 5 km southwest of Morrinsville. Five years' yield monitor data (2014, 2015, 2016, 2017, and 2018) were collected. The soil type is a mix of Te Puninga silt loam, some Ngakura loam and Morrinsville loam (Lilburne et al., 2012). The planting dates were missing for 2014 and 2015, and so from consulting the grower, the planting date for 2016 was assumed for those two years.

Site 4 (174.904 E, -37.314 S) is a 7.8-ha non-irrigated maize field located near Onewhero, approximately 10 km to the southern edge of the Auckland Region. The soil type is Karaka deep clay (Lilburne et al., 2012). Seven years' yield monitor data (2005, 2007, 2009, 2010, 2013, 2015, and 2017) were collected. The planting dates were missing for all years except for 2017. After consultation with the farmer, it was decided to use 10 October for planting for the missing years.

Site 5 (175.763 E, -37.798 S) is a 24-ha non-irrigated mixed arable field located less than 1 km north of Matamata. The soil type is mainly Te Puninga silt loam and some Ngakura loam (10 %) (Lilburne et al., 2012). Three years of maize yield data were collected (2008, 2009, and 2010).

To determine if a trained model derived elsewhere can predict yield from a maize field that has a limited dataset, the data from the five individual fields were pooled together to form a larger maize yield dataset to be analysed in the predictive modelling analysis and to identify important variables. All fields have soil EC, elevation and weather data collected (Table 3-2).

Table 3-2 Summary statistics (N = 2520) of the response variable yield and predictors for all five fields ("Rain.7" represents the accumulated rainfall within the week before planting; "Rain0" represents the accumulated rainfall in the first week (7 days) after planting; "Rain7" represents the accumulated rainfall in the second week (14 days) after planting; "rad" -- solar radiation; "GDD" -- growing degree days)

	mean	sd	median	min	max
yield	11.48	3.24	11.64	0.47	25.89
soilec_shallow	5.93	2.75	5.05	1.61	17.9
soilec_deep	5.61	2.81	5.38	-1.45	15.88
elevation	62.62	25.58	60.77	29.99	113.65
Rain.7	15.95	15.82	13	0	75.5
Rain0	20.82	14.49	16	0	44.5
Rain7	13.77	9.64	13	0.2	36.4
Rain14	22	16.33	18.6	0	58.5
Rain21	12.81	12.35	3.3	0	38.6
Rain28	13.09	12.6	6.7	0	46.4
Rain35	9.13	8.17	5.6	0	28.7
rad.7	121.01	13.84	123.1	100.76	145.12
rad0	126.04	15.73	121.75	88.49	155.93
rad7	140.78	19.94	143.39	101.52	171.55
rad14	138.4	20.77	139.92	102.85	174.9
rad21	148.46	16.37	151.59	116.92	175.51
rad28	149.43	13.04	151.06	121.11	183.41
rad35	154.75	19.75	156.45	98.65	195.3
GDD.7	37.57	6.89	35.85	22.35	51.95
GDD0	39.12	8.8	36.65	27.4	66.85
GDD7	41.37	11.53	39.1	26.75	77.2
GDD14	45.8	12.88	42.55	30.75	82.55
GDD21	45.84	13.02	38.95	32.6	76.9
GDD28	51.62	10.53	50.95	28.25	79.65
GDD35	56.74	10.61	59.25	35.65	74.1

3.4 Methods and approach to data analysis

3.4.1 Yield data pre-processing

Pre-processing of historical yield monitor data is required to reduce errors associated with the sensing systems and operations, and improve yield map quality (see section 2.3.1.1.1). Therefore, spatial data filtering was conducted to remove erroneous data values. Data filtering software is available to help analysts filter yield data to improve the accuracy of raw data. A freeware filter, "Yield Editor 2", developed by Sudduth et al. (2012) was initially used to establish its efficacy to remediate data errors. However, it was found that some historical yield datasets missed important attributes such as GPS timestamps, which led to processing difficulties in the software. To filter the spatial data with missing attributes, a tailored script was incorporated into R (R Core Team, version 3.5.3), which is available in Appendix 2.

Specifically, spatial data filtering is a two-stage process, which removes erroneous values sequentially. The erroneous values include:

- Data outliers: those values that are unrealistically large or small. Overly large yield values are commonly caused by abrupt stops of the harvesters, where the travel velocity reduces too abruptly while receiving the crop. The zero yield values are commonly located at field edges and headlands where the harvester is turning when it is likely that cutting stops and starts.
- Data inliers: Yield monitor data may contain errors where the yield values appear to be reasonable but are dissimilar to their neighbouring points due to measurement errors over short distances (described in section 2.3.1.1.1). To remove data inliers, the spatial relationships of the yield values need to be considered.

In this study, outliers were identified using boxplots, a simple univariate technique for detecting outliers and for presenting the distribution of data (Zhao & Cen, 2013). A boxplot is derived from the median value (or the middle value), as well as the lower- and upper- quartile values of the data. The upper quartile is the data value that has 25% of the population of values above it and 75% below. The lower quartile is the

data value that has 75% of the population of values above it and 25% below (King & Eckersley, 2019, pp.1-21). The data outliers were then determined based on the maximum MAX_Y and minimum MIN_Y thresholds derived from the median, and the lower- and upper- quartile values. Figure 3-7 demonstrates the procedure that was used for determining these thresholds.

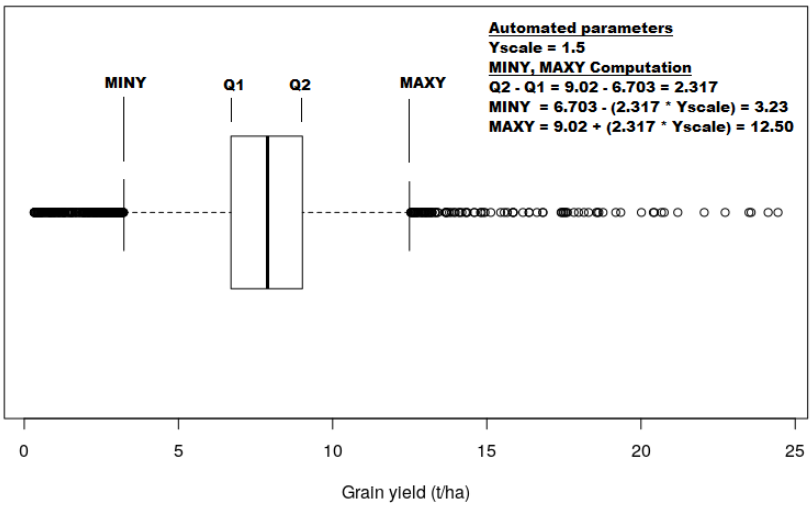


Figure 3-7 2017 harvest yield from NCRS boxplot and method for computing MIN_Y and MAX_Y filter parameter values. In this study, a scale factor $Y_{scale}$of 1.5 was arbitrarily selected. The greater the scale factor, the fewer points were labelled as the outliers and vice versa.

The upper $[Q_2]$ and lower $[Q_1]$ quantiles of the yield distribution were determined for the median value. The interquartile range Q_{inter} was then calculated by subtracting Q_2 from Q_1. To determine MIN_Y and MAX_Y, the upper and lower quantile values were expanded by a scale factor Y_{scale} of the interquartile range. This is expressed mathematically:

$$Q_{inter} = Q_2 - Q_1$$

$$MIN_Y = Q_1 - Q_{inter} \times Y_{scale} \qquad \text{(eq 3.1, 3.2, 3.3)}$$

$$MAX_Y = Q_2 - Q_{inter} \times Y_{scale}$$

Where:

Q_{inter}= interquartile range

MIN_Y=minimum threshold for yield

MAX_Y=maximum threshold for yield

Y_{scale}= scale factor for multiplication. Y_{scale}=1.5

To identify data inliers (values that are dissimilar from their neighbours), a generic spatial filtering technique developed by Spekken et al. (2013) was used. This technique requires spatial coordinates (longitude$X$ and latitude$Y$) and yield data values). Its implementation includes the following steps:

1. The coefficient of variation (CV) of a specific yield value Z_i (represented by the red dot in Figure 3-8) was first calculated in relation to the other points within a defined search window (5 m radius, selected arbitrarily based on the machine cut width to eliminate local variations produced during machinery operations while keeping the computation time of this analysis manageable) (Figure 3-8). The higher the CV, the more dissimilar the value of a specific point was compared to other points in the search window.

2. The points with CVs above 20% (arbitrarily selected based on Spekken et al. (2013)'s study) within a 5 m radius were identified.

3. If the number of yield values inside each search radius equalled the number of yield values with a high CV (>20%), that specific yield value (represented by the red dot in Figure 3-8) was labelled as an inlier and removed from the dataset.

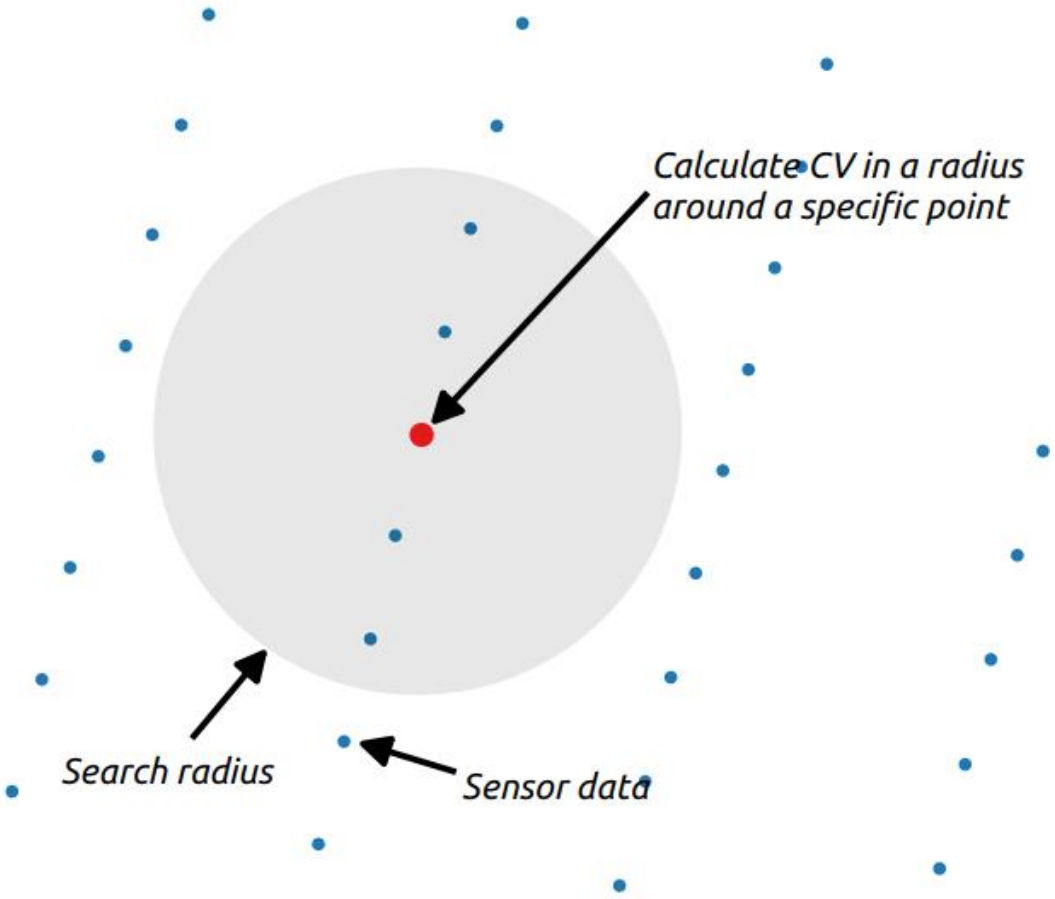


Figure 3-8 Inlier filtering process (a point was identified as an "inlier" based on the coefficient of variation [CV] within a defined search radius)

3.4.1.1 Evaluation of data filtering using variogram

To examine how well the spatial filtering technique could remove errors and improve the quality of mapping, a variogram analysis was conducted. Variograms and semivariograms are mathematical representations of Tobler's First Law of Geography (1970). Statistically, this law states that the variance between two points is dependent on their distance. Semivariances usually decrease with decreasing distance and increase as the separation distance increases. This is a premise for kriging interpolation, used to estimate values at un-sampled locations based on the spatial relationship derived from sampled locations. If two locations are overlapping each other (distance = 0), there should be no difference in their values. However, when a regression line is fitted to the semivariance values, it commonly intercepts the y-axis (distance = 0) at some non-zero value, known as the "nugget variance". In a typical variogram, nugget (C_0) often indicates measurement errors over distances less than the shortest sampling interval. If the nugget variance approximates the maximum semivariance, the kriging process is meaningless, as equal weighting would be given to each measured observation, implying that the predicted values are not related to the distances to nearby observation. Therefore, a reduction of the nugget variance of a variogram would indicate that the measurement errors over short distances are being remediated and thereby improving the outcome of the kriging process (Webster & Oliver, 1992).

In this study, the variogram was computed using the R package "automap". The package provides five different types of variogram models ("Sph" [spherical], "Exp" [exponential], "Gau" [gaussian], "Mat" [Matérn], "Ste" [Matern, M. Stein's parameterization]) (Hiemstra et al., 2009). The comparisons of kriging accuracy between before- and after-filtering were made for each variogram model using 10-fold cross-validation. These analyses evaluate the effect of spatial filtering on improving the accuracy of yield maps, which allows using kriging for modelling subfield yield variability.

3.4.2 Mapping spatial data using kriging

To integrate different data sets (yield monitor data, soil EC, soil OM, and elevation) which were originally surveyed at different locations within the NCRS field, spatial data points for each data set were interpolated into a raster map consisting of grid pixels.

The GPS coordinates of spatial data sets (longitude and latitude) were recorded using the standard World Geodetic System (WGS84) datum as decimal degrees, whereas kriging interpolation required the coordinates to be a distance unit (metre) to compute Euclidean distance. Therefore, spatial data sets mapped with the WGS84 datum were re-projected to the New Zealand Transverse Mercator projection (coordinate system documented by European Petroleum Survey Group [epsg]: code 2193): the coordinate reference system optimised for New Zealand. This transformation (using the R spatial analysis package "sp") facilitated the computation of distances between paired points (i.e. distances measured in straight lines), as required by kriging interpolation of yield and other spatial data such as soil EC. A regular grid of 6m × 6m cells was then created within the field boundary of NCRS. This cell size was selected because it was the application width of the maize seed planter.

Ordinary kriging is more suitable for interpolating yield data than simple kriging which assumes a constant mean yield over the entire field, and so does not reflect sub-field variations in the mean yield. The yield values at the grid nodes were interpolated based on nearby actual yield observations using ordinary kriging. Ordinary kriging is used to estimate local variations requiring only spatial X, Y coordinates, and the yield data values (Z). Ordinary kriging assumes a constant unknown mean in the local neighbourhood of each interpolated yield value. In this study, the premise is that the average yield is locally different from one neighbourhood value to another.

The principle of kriging is to predict the value at a given location by computing a weighted average of the known values nearby, which is mathematically closely related to regression analysis:

$$Z_0 = m_i + \sum_{i=1}^{N} \omega_i [z_i - m_i], \quad (3.4)$$

Where:

m_i= trend component (or drift), in simple kriging this is equal to the mean of z_i;

z_i = an observed value;

ω_i = the weight assigned to an observed value;

Let$\sum \ \omega_i = 1$ because the estimator is unbiased.

The purpose of kriging is to minimise the variance between predicted values and observed values (mean squared error, $\sigma_E^2 = E\left\{\left[Z_0 - \hat{Z}_0\right]^2\right\}$ by selecting ω_i in equation (3.4) based on the computed variogram. In ordinary kriging, a linear external parameter called the "Lagrange factor" (μ) is introduced to minimise the amount of kriging variance) and to improve prediction accuracy (Malvić et al., 2009).

In this study, the kriging variogram was computed using the R package "automap". This iterated through the entire list of variogram models ("Sph", "Exp", "Gau", "Mat", "Ste") and automatically fitted the model that produced a variogram with the smallest residual sum of squares (Hiemstra et al., 2009). The weights (equation 3.4) were then estimated based on the resulting variogram, providing interpolated yield values at un-sampled grid nodes.

Because there was a dense distribution of data points for producing adequate variograms, the ordinary kriging technique was applied to the yield data for each year to create yield maps, as well as for the generation of soil EC, soil OM, and elevation maps.

3.4.3 Delineating static zones for crop management

3.4.3.1 Yield-productivity zones

To present long-term yield variability in a measured and objective way and to inform soil sampling locations for soil texture analysis (section 3.3.5), yield-based MZs with high-yielding (HY) and low-yielding

(LY) zones were delineated. These were based on the interpolated yield maps (section 3.4.2) for the four harvests (2014, 2015, 2017, and 2018).

To produce static yield zones, a statistical technique proposed by Blackmore (2000) was used. The yield values from each season's yield maps were normalised against the average yield of that NCRS site (i.e. each yield was divided by field average which was set at 100%). A zone with a normalised yield higher than 100% was classified as `relatively high yielding' (HY) and vice versa (LY). To estimate how stable these yield values were, temporal variability was calculated based on the coefficient of variance (CV) at each grid node - the standard deviation of the mean over time. If over the four years, one zone sometimes yielded high and sometimes yielded low (relative to the mean), this would result in a high CV value, indicating temporal instability. Blackmore et al. (2003) noted that, in England, for most fields that grow grain crops, there are small areas that may be classified as "temporally unstable", which is consistent with mapping yield data from several maize fields (Holmes & Jiang, 2018). In this research, two types of yield zones were delineated (HY and LY). The delineation of these zones using a clustering method is described in the following section 3.4.3.2.

3.4.3.2 Soil zones (SZ)

To examine if soil EC maps can be used as yield predictors, soil zones were delineated based on the soil EC maps to establish that association with previous yield-productivity zones delineated. Fuzzy c-mean clustering (also known as soft k-means) was applied to the soil EC maps to generate a single map of different zones using the R software package ``e1071'' (Cannon et al., 1986).

Fuzzy c-means clustering has been embedded in several software packages (Fridgen et al., 2004; Minasny & McBratney, 2002) and is used extensively to partition data sets into groups to help delineate MZs (Kitchen et al., 2005; Lark & Stafford, 1997; Moral et al., 2010; Taylor et al., 2007). Fuzzy c-means clustering is an unsupervised classification algorithm (unsupervised classification is a term in machine learning used

to describe identifying patterns without any prior knowledge) and aims to minimise the sum of within-cluster variances through numerous iterations, expressed mathematically as:

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} u_{i,j}^{m} (x_i - \mu_j)^2, \tag{3.5}$$

Where:

x_i = the principal component scores

C_j = the centroid of a cluster and the mean of all points, is expressed as:

$$\frac{\sum_{x_i \in C_j} u_{i,j}^{m} x}{\sum_{x_i \in C_j} u_{i,j}^{m}} \tag{3.6}$$

$u_{i,j}$= the cluster membership$u_{i,j} \in [0,1]$, the degree to which an observation x_i belongs to a cluster c_j and $\sum_{j=1}^{k} \mu_{i,j} = 1$

μ_j= the centre of the cluster j

m= the fuzziness exponent, m is typically greater than 1. As m tends towards 1 it transitions from fuzzy to hard clustering, whereas a value of m tending towards infinity indicates complete fuzziness so that each observation has equal membership in all clusters.

3.4.3.3 Crop reflectance zones (CRZ)

To examine if satellite images can reflect the effect of spatial soil variability on yield, the crop reflectance images acquired from the satellites in section 3.3.3 were divided into two zones using fuzzy c-means clustering. The crop images (B, G, R, and NIR) are raster grids consisting of regularly spaced nodes, which are expressed as pixels.

For multi-band data, Principal Component Analysis (PCA) was performed. PCA is a method of eliminating features by undertaking orthogonal transformation (rotating the coordinate systems to find the maximum

variance of the data). The results of PCA may show that the original data can be compressed into new variables (coordinates) called Principal Components (PCs). The first two PCs generally explain most of the variation in a multivariate dataset and are ranked in order of influence on data variation.

The results of PCA are often discussed in terms of:

- Loadings - the weight by which each standardised original variable should be multiplied to obtain the scores (i.e. the transformed values). Loadings represent how strongly each original variable contributes to the transformed PC. Variables with (+) or (-) loadings indicate that these variables both positively influence (+) or negatively influence (-) the PC.
- Eigenvalues - the amount of variation retained by each PC.

One necessary task with the clustering technique is determining the number of clusters. Instead of choosing an arbitrary number, the optimal number of clusters was selected using the silhouette method (Rousseeuw, 1987). This method measures how well the points are clustered into similar groups by evaluating the average distance between within-cluster points against the average distance between separate-cluster points. This measure produces a range between -1 and 1. An average silhouette near 1 indicates that the samples are far away from neighbouring clusters and the groups are more isolated. A value of 0 indicates that the samples are on, or very close to, the boundary between two neighbouring clusters. A negative value indicates that the samples might have been misclassified. An average silhouette is equal to 0 if only one cluster is produced. Figure 3-9 shows that the optimal number of clusters was determined by the highest average silhouette width.

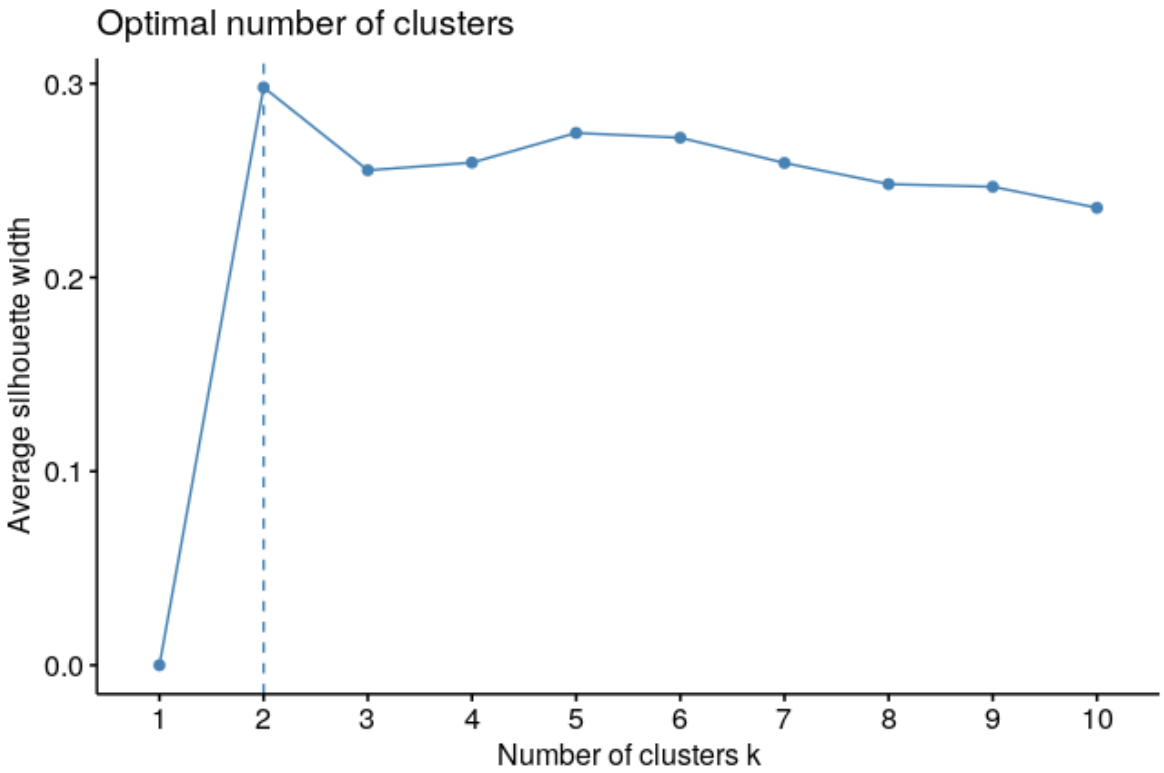


Figure 3-9 The average silhouette was computed for every increase in the number of clusters. The highest silhouette was produced when two clusters were created.

3.4.3.4 Comparison of static zones

To determine if there is any association between yield productivity zones, crop reflectance zones and soil zones, the yield productivity zone map was compared with both crop reflectance zones and soil zones using areal agreement (the percentage of pixels consistently classified as HY and LY), and Cohen's kappa coefficients (how likely it is that the areal agreement occurs by chance). The kappa coefficient (κ) is mathematically expressed as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{3.7}$$

Where p_0 is the observed agreement amongst the zone maps, and p_e is the hypothetical probability of a chance agreement. A kappa of 1 indicates perfect agreement, whereas a kappa of 0 indicates agreement equivalent to chance (Landis & Koch, 1977). The results of this analysis provide an understanding as to

whether the spatial pattern of crop reflectance and long-term yield productivity is related to soil variability within-field and thus inform soil sampling locations (see section 3.3.5).

3.4.4 Soil texture

To calibrate soil EC data and further validate the effect of soil texture, soil samples collected from the NCRS site (section 3.3.5) were kept moist in a chiller, organic matter removed from subsamples by boiling with hydrogen peroxide (35% strength), and then analysed for texture using standard pipette sampling from settling suspensions (Claydon, 1989). Particle size classes were sand > 63 microns, silt 2 - 63 microns, and clay < 2 microns. To determine the precision of this procedure, triplicate analyses were conducted at selected soil depths.

3.4.5 Multivariate modelling analysis

To demonstrate the viability of estimating potential yield within-field, this study investigated the use of supervised machine learning models. The modelled output can then help inform crop management input decisions. Figure 3-10 illustrates the workflow for the multivariate analysis, which included the following aspects:

1. Model optimisation and hyperparameter tuning to avoid overfitting of the model, i.e. to avoid the model predicting well in training but predicts poorly in the validation.
2. Computing variable importance and variable elimination to further eliminate statistically insignificant variables to simplify the model, reducing computational time and providing interpretation as to how yield responded to the important spatiotemporal variables.
3. Evaluation of model performance for predicting yield

a. Multiple-year analysis to extract the relationship between yield and temporal factors using yield data from all available years.

b. Leave-out-one-year analysis to evaluate how the models predict the unseen data, i.e. the yield of the left-out year.

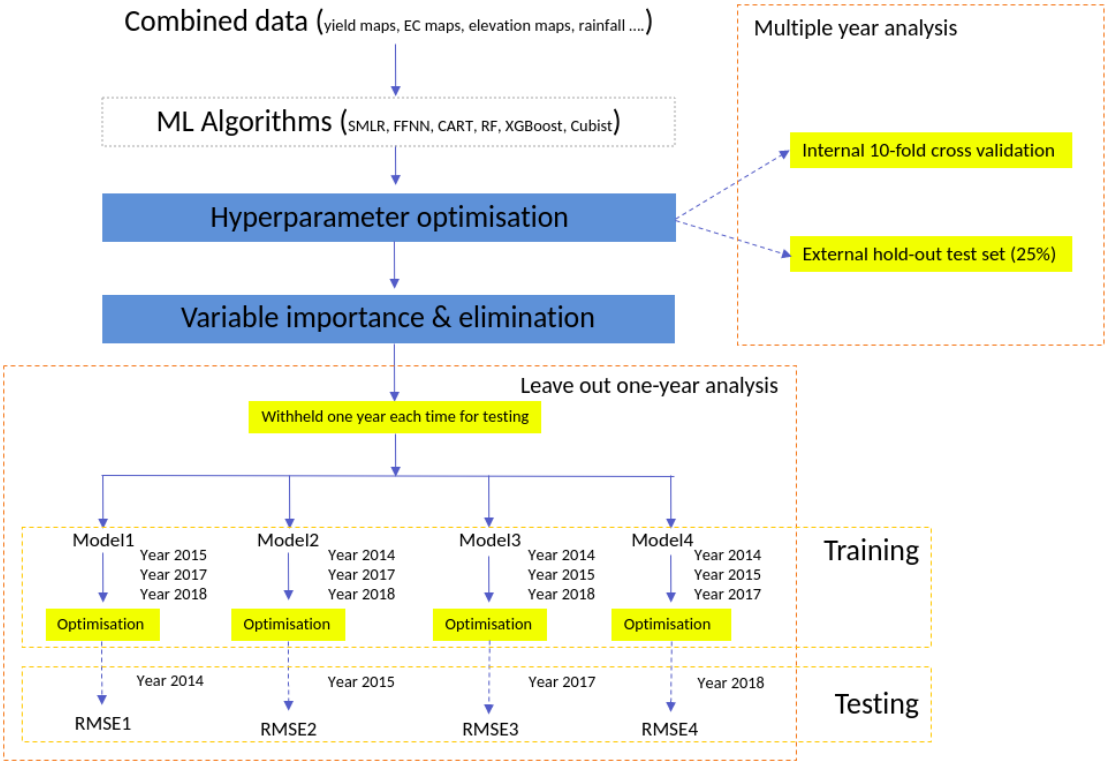


Figure 3-10 Flowchart of multivariable modelling analysis

The multivariate modelling analysis was conducted using R statistical software (R Core Team, version 3.5.3). Under the R machine learning library "caret" developed by Kuhn (2008), several packages were implemented for different model algorithms (Table 3-3).

Table 3-3 Algorithms, R packages and key principle

Algorithms	R packages	Key principle
Stepwise multiple linear regression (SMLR)	Built-in	Least-squares
Feed-forward neural network (FFNN)	nnet (version 7.3–12)	Gradient descent backpropagation
Classification and regression tree (CART)	rpart (version 4.1-11)	binary splits
Random forest (RF)	randomForest (version 4.6-12)	Bagging
Extreme gradient boosting (XGBoost)	xgboost (version 0.90.0.2)	Boosting
Cubist	Cubist (version 0.2.2)	Boosting

Multiple linear regression is the most common algorithm used to model the relationship between one response variable (yield) and many predictors, by using linear regression parameters. The model fits a line (or a plane) using the "least squares" method, i.e. the minimum sum of squared residuals and estimates all the β_i coefficients. The intercept ε can be included to minimise the error in the fit. This multiple linear regression is expressed mathematically in equation 3.8:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon, \quad \text{(eq 3.8)}$$

Where:

β_i= linear coefficients;

X_i = the predictors, e.g. soil EC, elevation, OM, rainfall, solar radiation, GDD (see Table 3-2);

Y = the response variable, i.e. maize-grain yield;

In this study, multiple linear regression was undertaken using R statistical software. Multiple linear regression aims to measure the correlations between predictors and the response variable. Multicollinearity arises when two predictors are highly correlated with each other, which prevents the model from revealing the underlying relationship between a predictor and yield. Multicollinearity was first eliminated by filtering out the absolute pairwise linear correlation coefficients (Kuhn, 2008) between

predictors that were above a predefined threshold (0.8), which is widely used to indicate a statistically significant correlation (Mukaka, 2012; Buytaert et al., 2006). Predictors were then manually eliminated one-by-one, based on their statistical significance, with higher *p*-values ($p > 0.05$) eliminated first. The elimination process was repeated until no further variables could be deleted without a statistically significant loss of model accuracy (indicated by the coefficient of determination R^2). This process is called "stepwise multiple linear regression" (SMLR). The results from stepwise multiple linear regression can help to identify the cause-and-effect relationships between maize-grain yield and the predictors such as soil EC and elevation.

The stepwise multiple linear regression model assumes the relationships between the response variable (maize yield) and the predictors to be linear. To explore any non-linear relationships, feed-forward neural network (FFNN) models have been used by previous studies (Drummond et al., 2003). To be able to compare this study with their findings, feed-forward neural network modelling is also implemented using the "nnet" package (Ripley et al., 2016) in the R statistical programme.

A feed-forward neural network is a simple type of neural network with only one hidden layer (Figure 3-11). The input predictor data $X[x_1, x_2, ... x_n]$ moves from the input- to hidden-layer and then to the output layer (predicted yield) through fully connected nodes in a single direction (Figure 3-11). The nodes in the hidden layers are known as "neurons" and each neuron contains a non-linear "activation" function. Each neuron receives the weighted average of the predictor data $\Sigma[w_{11}x_1, w_{12}x_1, ... w_{1n}x_1]/n$ and then transforms the data into a sigmoid curve with each output value ranged from 0 to 1. Then based on linear regression of the secondary data from the hidden layer, the final output value is estimated. The weights $W[w_{11}, w_{12}, ... w_{nn}]$ and $G[g_1, g_2 ... g_n]$ in the network are randomly initiated at the start and are updated by iteratively evaluating the derivatives of the "cost" function (sum of squared error). The optimisation of the network is completed when the "cost" reaches the minimum. This optimisation procedure is known as the "gradient descent", and is designed to reduce the computational complexity of the neural network (LeCun et al., 1998).

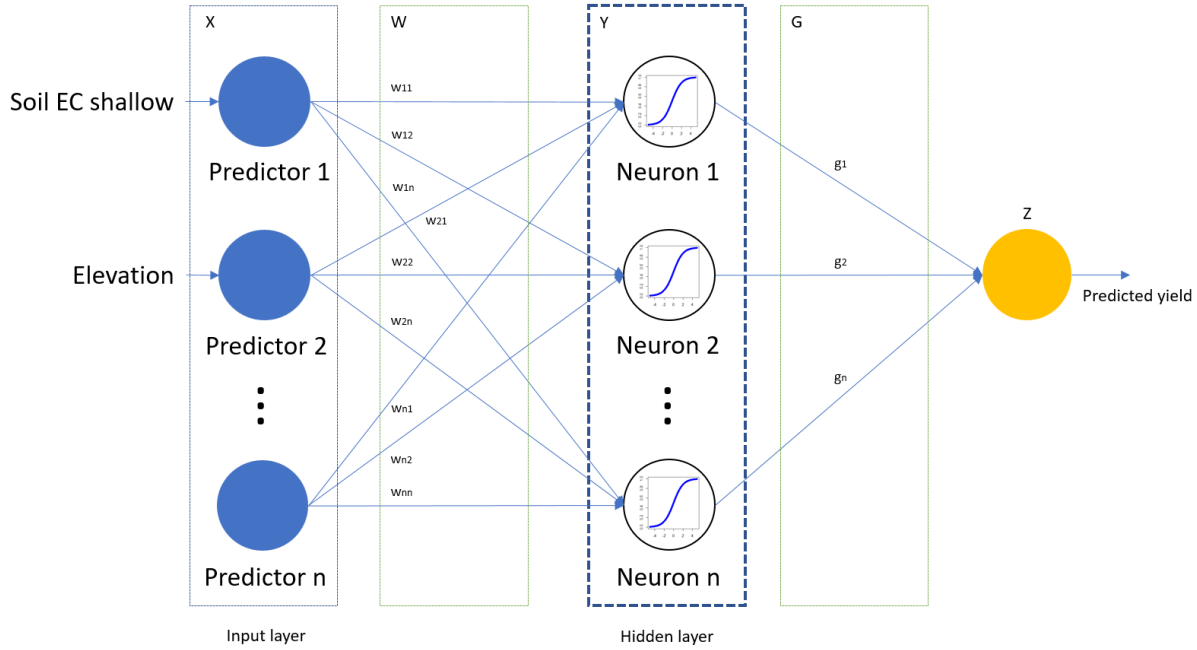


Figure 3-11 Feedforward neural network structure (X represents the matrix of input data; W and G are the weight matrices assigned between the layers, which are updated iteratively to achieve the lowest prediction error; Y and Z are the matrices of the output values for that layer)

In contrast to multiple linear regression and neuron networks, the classification and regression tree (CART) was originally developed to solve the interaction effects (the effect of one predictor on the response variable, depending on the state of another predictor), and to identify multicollinearity in linear regression, with inputs from a mix of continuous and categorical data (Morgan & Sonquist, 1963). Given the mix of spatiotemporal predictors employed for multivariate analysis in this study, the CART model may be applicable for predicting yield because it implicitly eliminates unimportant predictors from the training process (Figure 3-12). CART aims to partition data into homogeneous subsets via a top-down and recursive process (a regression tree):

- CART iterates through each observation of a predictor and computes the sum of squared errors (SSE) from their true value and the average value of that partitioned subgroup. The observation value with the lowest SSE will be used to make an initial split to determine the threshold for the prediction.

- The algorithm then computes the SSE for each predictor. The variable producing the lowest SSE is used as the first “root node” at the top of the tree.
- The process is then repeated until a certain tree size threshold is met.

The CART model was implemented using the package ``rpart'' (Therneau & Atkinson, 1997). The tree size threshold in this model was defined as the maximum tree depth (“maxdepth”). A tree that is too small may produce a large yield prediction error because only a few predicted values are produced (Figure 3-12). A tree too large will produce too many splits (decision points). Therefore, the “maxdepth” parameter was tuned using a grid search method (see section 3.4.5.1). CART often provides less accurate results due to binary splits and classed outputs (Ließ et al., 2012). To enhance the prediction ability of a single CART tree, ensemble tree methods are also evaluated in this study, including Bagging, Boosting, and Cubist Regression.

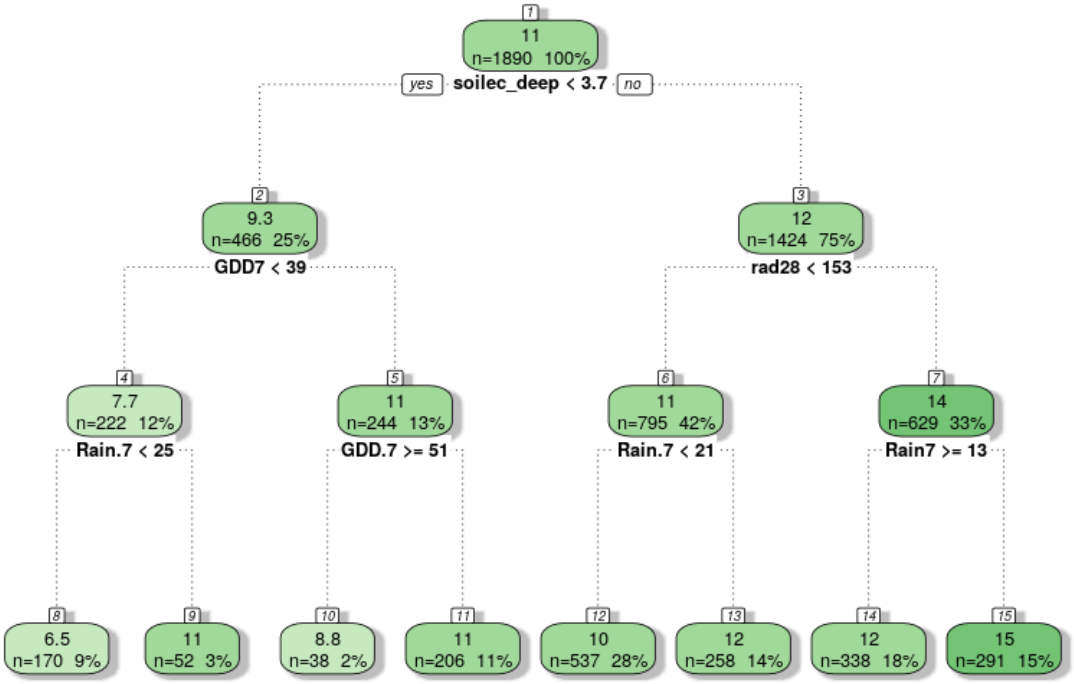


Figure 3-12 CART Decision tree (Data was split into two subgroups from layer to layer. In this model, the predictions of yield were then made by a series of constraints defined by the input predictors and their splits (decision points). The data that matched the condition would go to the left-side branch, i.e. a value of the splitting predictor was greater than or, equal to, a statistically defined threshold. Otherwise, the data would go to the right-side branch if the value of the splitting predictor was less than the defined threshold)

Bagging, known as "random forest", is a collection of regression trees created using "bootstrapped" datasets and uses all these trees to predict the response variable. Bootstrapped datasets are new datasets created by sub-sampling (random sampling with replacement) the original dataset. The observations that have not been sampled is called "out-of-bag" samples and are used to evaluate the prediction error in random forest processing. Several predictors are then randomly selected to build a new tree based on the "bootstrapped" dataset. As a result of random forest processing, many trees are built and the final prediction of yield is the average value of the yield values of all tree predictions (Breiman, 2001). The random forest model was implemented using the "randomForest" package in R (Liaw & Wiener, 2018).

Boosting creates trees sequentially, based on the prediction produced by the previous tree created. The boosting model then updates the residuals of prediction by creating new sequential trees until the model converges to the stage where the residuals of prediction can no longer be reduced. The magnitude of each update is controlled by a parameter called "learning rate" (η). It is the step weights given to the prediction input of each tree to ensure convergence. XGBoost (Extreme Gradient Boosting) is a more complex sequential boosting algorithm, originating from Breiman (2001)'s paper. XGBoost introduces a regularisation term (λ) which is applied to each tree to avoid overfitting. A larger λ ($\lambda > 0$) causes more pruning of each tree and produces smaller sums of residual and therefore better prediction. XGBoost is a model that has recently been dominating applied machine learning for structured or tabular data (Nielsen, 2016). The XGBoost model was implemented using the "xgboost" package in R (Chen et al, 2015).

Cubist regression has been used for some large remote sensing datasets and reported promising prediction accuracies and less computation time compared to that of random forests (Noi et al., 2017; Zhou et al., 2019). However, it is a rule-based decision tree algorithm introduced to the R statistical program (Kuhn et al., 2012). It is a proprietary product and thus little publicly available documentation exists as to how the model functions. The application of this model is, however, popular and has been cited in literature since it was introduced into R in 2012. Distinct from regression trees such as CART, where values are predicted at their terminal nodes, the Cubist model produces a set of rules ("if-then" statements), with each rule comprising multivariate linear regression models at the terminal nodes (Table 3-4). The Cubist model adds a boosting procedure called "committees", which are a series of trees created sequentially. The Cubist model adds a refinement to the boosting procedure by using neighbour–based adjustments so that each predicted value can be adjusted based on observed values in a similar neighbour cluster determined from the training set (Walton, 2008). The Cubist regression was undertaken using the "Cubist" package in R.

Table 3-4 Example of Cubist modelling output (first 10 rules)

Rule	Condition	Linear regression model	support	mean	error
1	elevation > 38.35374 & elevation <= 39.03139 & rad14 > 144.61 & rad35 > 156.45	(3.3902982) + (0.49 * soilec_deep)	27	4.4	0.9
2	soilec_deep <= 4.073236 & rad14 > 124.1 & rad14 <= 144.61	(-22.5182272) + (0.132 * rad14) + (0.39 * soilec_deep) + (0.047 * rad21) + (0.028 * elevation) + (0.048 * GDD21) + (0.015 * GDD14) + (0.007 * rad0) - (0.03 * soilec_shallow) - (0.01 * GDD.7)	92	7.5	1.4
3	soilec_deep <= 4.824256 & elevation <= 38.35374 & rad14 > 144.61 & rad35 > 156.45	(3.7557431) + (1.82 * soilec_deep) - (0.007 * rad35) + (0.004 * rad14) - (0.03 * soilec_shallow)	21	8.4	3.8
4	soilec_deep <= 3.612189 & elevation <= 38.1531	(-20.1481427) + (3.84 * soilec_shallow) + (0.049 * elevation) + (0.016 * rad14) + (0.017 * rad35) + (0.013 * rad28) + (0.05 * soilec_deep)	38	9.3	3.0
5	soilec_deep > 4.824256 & elevation <= 35.46135 & rad35 > 156.45	(5.560037) + (0.36 * soilec_shallow)	46	9.5	2.4
6	soilec_shallow > 5.35793 & rad14 <= 144.61 & rad35 > 156.45 & GDD14 > 36.65	(316.3637189) - (2.185 * rad35) + (0.524 * elevation) + (0.241 * rad14) + (0.89 * soilec_deep) - (0.6 * soilec_shallow)	110	10.2	2.0
7	rad14 <= 144.61 & rad35 <= 156.45 & GDD14 > 36.65	(14.6920276) - (0.046 * elevation)	131	10.9	1.1
8	soilec_deep > 3.612189 & elevation <= 38.1531 & rad28 <= 166.29 & rad35 <= 156.45	(-62.1986031) + (0.424 * elevation) + (0.255 * rad14) + (0.063 * rad35) + (0.19 * soilec_deep) + (0.03 * rad28) - (0.07 * soilec_shallow)	147	11.0	1.8
9	elevation > 38.1531 & rad14 > 144.61 & rad35 <= 156.45	(25.847636) + (0.258 * rad35) + (0.18 * elevation) - (0.331 * rad28) - (0.055 * rad14) + (0.4 * soilec_deep) - (0.02 * soilec_shallow)	157	11.2	1.3
10	elevation <= 35.81275 & rad28 > 166.29	(-20.9991053) + (0.84 * soilec_deep) + (0.074 * elevation) + (0.056 * rad14) + (0.057 * rad35) + (0.028 * rad28) - (0.07 * soilec_shallow)	30	11.2	4.1

3.4.5.1 Optimising model hyperparameters

Hyperparameters are the parameters of a model used to control the model learning process such as its complexity or how quickly it should learn. To ensure the best performance of the prediction models and to avoid overfitting (i.e. model predicts well for the training set but predicts poorly for the test set), the models were optimised using a grid search method with *K*-fold cross-validation. The "cost" function such as root mean squared error (RMSE) was computed for each combination of parameters making up each axis of the grid (or grids). The lowest RMSE was then determined for a specific set of hyperparameters (Chan & Treleavan, 2015).

To avoid bias in data selection, a 10-fold cross-validation method was used for hyperparameter tuning. Although there is no strict rule for determining the number of folds (*K*), a value of $K = 5$ or 10 may produce less bias prediction in the field of applied machine learning (Rodriguez et al., 2009). In this study, $K = 10$ was selected for hyperparameter tuning considering enough samples for training the model and potentially large spatiotemporal yield variability in the multiple-year yield data. $K = 5$ may not be enough to capture that variability whereas $K > 10$ would increase computation time. In this method, the data were randomly divided into 10 subsets of equal size, in which one subset (10% of the total dataset) was used as the validation subset, and the remaining data (90%) as the training subset to train a model with a set of hyperparameters. The validation subset was used to evaluate the performance of the trained model and computed the RMSE for each model (Table 3-5). This cross-validation process was iterated and the overall RMSE was calculated as the average value of all modelled RMSEs for each set of hyperparameters. RMSEs from different sets of hyperparameters were compared. The hyperparameters that resulted in the smallest overall RMSE for a model were selected as the optimum for that model. Table 3-5 summarises the use of 10-fold cross-validation to evaluate the effect of the combination of the hyperparameter (a) and hyperparameter (b) on the model performance, indicated by the RMSEs.

Table 3-5 Illustration of the 10-fold cross-validation method for hyperparameter optimisation

Hyperparameters	Training set	Testing set	Performance indicator
a, b	Fold1 + Fold2 + … + Fold9 (90%)	Fold10 (10%)	RMSE1
a, b	Fold2 + Fold3 + … + Fold10 (90%)	Fold1 (10%)	RMSE2
a, b	Fold1 + Fold3 + … + Fold10 (90%)	Fold2 (10%)	RMSE3
…	…	…	....

For a single linear regression model, a regression line is fitted to minimise the sum of squared errors. In an ideal situation, without interference from other variables, the regression line should go through the origin of the X-Y axes. However, in applied situations, the regression line always intercepts the Y-axis at some non-zero "nugget" value. This intercept is then interpreted as the expected mean value of Y when X is zero. Since yield cannot be negative, a range of integer values for the intercept from 0 to 10 was tested in 10-fold cross-validation.

For feed-forward neural network (FFNN) modelling, two hyperparameters were tuned, including the number of neurons in the hidden layer (hidden units) and weight decay. The number of neurons in the hidden layer (hidden units) determines a FFNN model's ability to minimise prediction RMSEs for each subset of the training data. The optimal number of hidden units is often a value between the number of nodes in the input layer and that of the output layer (Heaton, 2008, p.159). Too many hidden units may cause overfitting, whereas too few hidden units decrease the complexity of the network. The neural network model implemented by the "nnet" package also adds a regularisation term called "weight decay" to the gradient descent algorithm by further regulating the size of the step in the descending process to prevent overfitting. Models with several hidden units ranging from 1 to 5, in combination to a weight decay value of 0.1, 0.2, and 0.3, were tested and is a suitable weight decay range for tuning the model hyperparameters in a grid search with manageable computation time. A larger weight decay value could result in a greater decrease in RMSE while there were only a few hidden units (functionally like regression).

As the number of hidden units increased, the prediction RMSE may become insensitive to the weight decay value.

For classification and regression tree (CART) modelling, a prediction is made by repeating binary splits of the data until a predefined limit to the number of tree layers is met. In this analysis, several tree layers (ranging from 1 to 20) was tested in the grid search. For regression trees, fewer layers in the tree mean that fewer values are less able to be predicted, resulting in poor model performance. However, if the tree has more layers then model performance could be compromised due to overfitting.

For random forest (RF), a yield prediction is made by combining the predicted values of each simultaneously created CART tree. Each tree is created using a "bootstrapped" (random sampling with replacement) dataset randomly sampled from the original dataset. The number of predictors that are randomly selected ("mtry") determines the structural variation between any two trees in the forest. Increasing the number of predictors randomly selected may improve the strength of each tree, but decrease the variation between individual trees, causing similar trees and similar predicted values that could bias the final prediction. The cross-validation models with several randomly selected predictors ("mtry") were tested, ranging from one predictor to the maximum number of predictors (24) for hyperparameter tuning and finding the lowest RMSE.

XGBoost is a complex model based on "gradient boosted trees", where trees are created in sequential order, with each tree created based on the prediction made by the previous tree created. XGBoost also integrates a regularisation term when building trees for pruning, as well as providing options for adjusting the tree size, learning rate, and observation and variable subsampling to reduce bias. Many hyperparameters may influence model performance. The important ones are tree depth "Max Tree Depth", learning rate "eta", subsample ratio of the training instances "subsample", subsample ratio of columns "colsample_bytree" and the number of boosting iterations. As with a single CART tree, the "Max Tree Depth" controls the number of layers of each tree. The learning rate "eta" (η) determines the step weights given to the prediction input of each tree. Increasing η speeds computation due to fewer

iterations (and trees), but does not ensure the lowest RMSE. A smaller η reduces computation speed since more trees are added to the model to ensure convergence. Therefore, in this study, small η values in the range of 0.1 to 0.4 were tested for hyperparameter tuning. The hyperparameters "subsample" and "colsample_bytree" determine the sampling ratios of the dataset for building the trees. An increased number of iterations lowers residual errors so that, over time, the boosted trees should achieve the lowest RMSE possible.

The Cubist model is a proprietary product and has very little public documentation to support its use. Based on limited information, it appears that the Cubist model uses a boosting-like procedure called "committees". As with the XGBoost model, increasing the number of "committees" (the number of boosting iterations) may help reduce prediction error and prevent overfitting. Also, the Cubist model introduces neighbour–based adjustments, so that each predicted value can be adjusted based on observed values in a similar neighbour cluster determined from the training set. In this analysis, four committee values (#Committees - 1, 5, 10 and 20) were tested in a grid search. The Cubist model is computationally efficient and processes large multi-temporal images in remote sensing relatively quickly, making it a potential candidate for predicting spatial yield.

3.4.5.2 Assessing the model performance

3.4.5.2.1 Multiple year analysis

To select the best model for predicting yield, an internal "split-sample" approach was used to measure prediction accuracy. In this approach, the dataset was randomly partitioned into two data subsets:

1. A subset consisting of 75% of the data was partitioned as the "training set" for fitting and training the model for prediction, and
2. An independent subset consisting of the remaining 25% of the data is withheld as the "test set" and used to test the accuracy of the model.

In the multiple-year analysis, data subsets were created from the yield data for all available years. The prediction model was then constructed and used to predict the yield in the test set. The level of accuracy for a test model was calculated as the root mean squared error (RMSE) and R-squared (R^2), representing how well the predicted yield values are close to the actual yield.

3.4.5.2.2 Leave-out-one-year analysis using individual-field data

A trained model that successfully captures multivariate relationships should have the ability to predict independent field data. Since a relatively small number of years' yield data are available for all available sites, the multiple year data was cross-validated by year.

In cross-validation by year, one year of data was withheld as a test set for each iteration, with all remaining years included in the training set. The training set was used to predict yields for the year that was excluded as a test set (Table 3-6). This process was iterated for all four years and RMSEs were calculated. This indicated the ability of the trained model to handle new information, being the yield data collected from an additional harvest from another year.

Table 3-6 Datasets used in the leave-out-one-year analysis (FAR's NCRS).

Model	Training set	Test set
1	2014, 2015, 2017	2018
2	2014, 2015, 2018	2017
3	2014, 2017, 2018	2015
4	2015, 2017, 2018	2014

The results of this analysis will demonstrate the ability of the model to predict the yield of additional harvests in the same field, and the viability of delineating dynamic crop MZs based on predicted static

yield maps, which may lead to a more accurate prescription for crop management inputs such as N fertiliser application, compared to static yield maps (see section 3.4.3).

3.4.5.2.3 Leave-out-one-year analysis using pooled data from other sites

To determine if the trained model can predict yield from a maize field that has a limited dataset, the leave-out-one-year analysis was undertaken using data pooled from other maize field sites. This analysis will help determine whether the models trained for one field can be used to determine yield for other independent non-irrigated maize fields.

3.4.5.3 Analysis of predictor importance on pooled data

To identify important spatiotemporal factors and potentially provide insights into managing spatial yield variability, the data collected for four other non-irrigated maize fields and years were pooled together, and the relative importance of each predictor was computed for each trained model using the "varImp" function in the R "caret" library. Table 3-7 summarises the key indicator for computing predictor importance for each model.

Table 3-7 Key parameters for computing relative predictor importance for each model.

Algorithms	Key indicator
SMLR	Absolute t-value of a predictor
FFNN	The sum of the absolute weights connecting to a predictor
CART	Sum of information gained from the splits of a predictor
RF	Reduction of the prediction accuracy for "out-of-bag" samples after re-shuffling the
XGBoost	Number of times a predictor used for splits, divided by the number of trees created
Cubist	Percentage of a predictor used in the "if-then" conditions and the regression models

For SMLR, the importance of a predictor is determined by using its absolute t-value which indicates its statistical significance (p-value). The larger the absolute t-value, the smaller the p-value and more likely the null hypothesis will be rejected i.e. the yield is independent of or not linearly related to the predictor (Greenwell et al., 2018).

For FFNN, the importance of a predictor is estimated based on the sum of the absolute value of the weights $\Sigma[|w_{11}|, |w_{12}|, ... |w_{1n}|]$ connecting the predictor (Figure 3-11). The sum of absolute weights is then scaled from 0 to 1 relative to the most dominant predictor which is given the value of 1 (Garson, 1991).

For the CART model, each split results in a reduction of the mean squared error, because more similar subgroups are created. The importance of a predictor is calculated as the sum of the reduction of the prediction error attributed to each predictor at each split (Therneau & Atkinson, 1997). Therefore, the higher the reduction of the error, the higher importance the predictor has in determining the split.

For the RF model, the observations randomly sampled are used to create a "bootstrapped" dataset and build individual trees. The remaining observations that have not been used to build trees are called "out-of-bag" samples. The predictor importance is evaluated by predicting "out-of-bag" samples using the trained model. After re-shuffling the predictor of interest, the predictor importance is calculated to determine how much prediction accuracy would deteriorate. The re-shuffling breaks the relationship

between the predictor and the response variable (maize yield). If the accuracy of prediction reduces significantly after re-shuffling a predictor, the predictor is then considered important (Breiman, 2017; Fisher et al., 2018).

For the XGBoost model, predictor importance is calculated on the number of times a predictor is used to generate a split in each of the boosted trees. This is then divided by the total number of trees created (Hastie et al., 2009, p. 367).

For the Cubist model, predictor importance is calculated as the percentage of predictor usage in both the conditions and the linear regression models in the terminal nodes.

3.5 Summary

This chapter describes the study location, data, and analysis methods required for this research, which contains the following tasks:

- To improve the accuracy of yield maps in subsequent analyses, a customised spatial filtering algorithm was developed and tested by comparing modelled variograms and comparing the predicted values and the observed values in 10-fold cross-validation.
- To identify appropriate yield predictors, yield monitor data collected from NCRS were mapped and delineated into MZs, which were correlated with MZs derived from soil electrical conductivity (EC), soil organic matter (OM), and elevation, and MZs derived from multispectral crop images. Zonal soil sampling and soil texture analysis were undertaken to attempt to explain the yield variability and to provide a calibration for the use of soil EC.
- To determine the viability of predicting spatial yield using data that is readily available and inexpensive, in addition to NCRS, data were also collected from four other non-irrigated maize fields in Waikato. Temporal data (rainfall, solar radiation, and GDD) at various growth periods were collected from a nearby weather station and were incorporated into the development of subfield yield prediction models. Several machine learning models (SMLR, FFNN, CART, RF, XGBoost, and Cubist) were then implemented to predict yield, with the prediction errors of each model evaluated using “split-sample” approaches. To provide insights into managing subfield yield variability and the opportunity to fine-tune the prediction models, predictor importance analyses were undertaken using the pooled data from all available fields.

Chapter 4 Results

4.1 Introduction

To address the objectives outlined previously in Chapter 1, this chapter presents the results from the following analyses:

- Pre-processing of yield monitor data. Pre-processing was applied to all yield monitor data available for the study area because of the need for accurate maps on which to base site-specific management (section 4.2);
- Identifying spatiotemporal yield predictors by examining spatial and temporal variability of yield and soil. This includes mapping of spatial data (yield monitor data, soil EC, elevation, soil OM); delineating potential MZs and; examination of soil subfield and depth variability, in an attempt to explain the underlying reason for crop and yield variability (section 4.3);
- Prediction of yield using supervised machine learning models. The purpose of multivariate modelling analysis was to see if it is viable to delineate dynamic MZs that are adaptable to the effects of short-term temporal conditions (section 4.4).

4.2 Pre-processing of yield monitor data

To improve the quality of yield maps that could be produced from available yield monitor data, spatial filtering was conducted using the customised programme (Appendix 2) to eliminate yield data errors before kriging interpolation (see section 2.3.1.1).

The filtering programme targeted outliers (extremely high and low values that skew data distribution) and inliers (values that are dissimilar to the neighbours within a distance defined by the user, i.e. 5 m) for removal. Between 6 to 40% were removed from the original yield datasets (Table 4-1). The yield data for most years (2015, 2017, and 2018) had less than 1,000 points (6 to 12%) removed. The greatest number of points removed (40%) was associated with the yield in 2014.

Table 4-1 Statistics of yield monitor data values before and after-spatial filtering.

Year	mean		sd		min		max		obs		
	t/ha										
	before	after	before	after	before	after	before	after	before	after	Removed(%)
2014	7.1	8.5	6.3	3.6	0	1.1	98.0	19.7	11452	6871	40
2015	13.4	13.5	8.3	3.2	0	3.2	101.5	23.0	6333	5555	12
2017	7.8	7.9	2.1	1.6	0.3	3.2	24.4	12.5	9017	8431	6
2018	11.0	11.2	2.7	1.8	0.4	5.8	25.0	16.4	8592	7773	10
All year	-	10.0	-	3.4	-	1.1	-	23.0	35394	28630	19

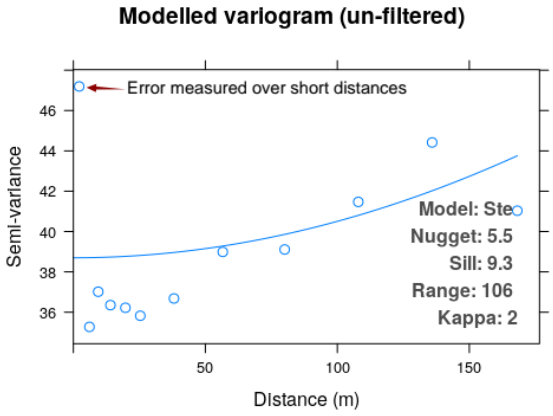
Geostatistical variogram analysis was conducted to evaluate the filtering performance. Figure 4-1 presents the comparisons of the variograms between using the unfiltered and filtered yield data for each year. These variograms were fitted with models that generated the smallest sum of squared errors for the aggregated semi-variance values (Hiemstra et al., 2009).

In Figure 4-1, the sill variance is the maximum semi-variance modelled between the values of paired locations. The nugget variance is the semi-variance when the distance that separates paired locations is at **0** (i.e. two locations overlap and hypothetically, they should have no difference in value). The nugget variance represents unnatural variations occurred over the sampled interval, which is commonly related to the measurement errors. A nugget/sill ratio close to 1 indicates a spatially random pattern at a specified resolution, which violates Tobler's First Law of Geography (1970) for natural variations and the premise for modelling the spatial structure of the data and kriging (Krige, 1951) (see Method section 3.4.1).

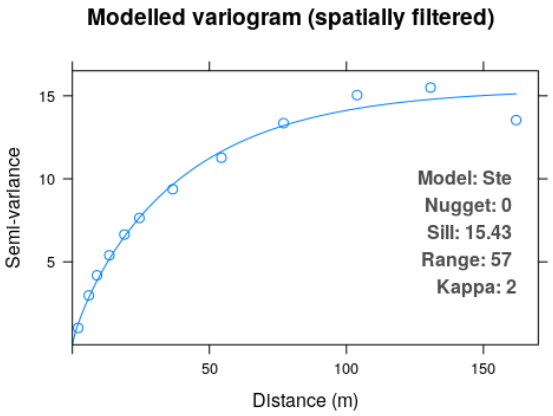
The variograms based on the un-filtered yield data had higher modelled nugget variances (Figure 4-1a, c, e, g), caused by measurement errors at short distances. These nugget variances were reduced to 0 on the modelled variograms, based on the spatially filtered yield data (Figure 4-1b, d, f, h). The results (Figure

4-1) showed that the applied spatial filtering method reduced the nugget/sill ratio from 0.18 – 0.91 to 0 for all available years (2014, 2015, 2017, and 2018).

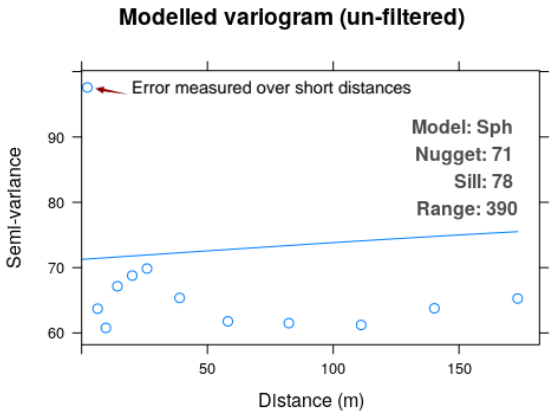
Figure 4-1c showed a high modelled nugget variance (71) and that an increase of distance separating paired locations had little effect on their semi-variances. After spatial filtering, the semi-variances increased initially as the distances increased (Figure 4-1d). The curve plateaued as it reached a threshold distance, meaning that the semi-variances were no longer dependent on the separation distances. The reduction of nugget variance suggested a successful elimination of the yield sensor errors for modelling yield variations over short distances.



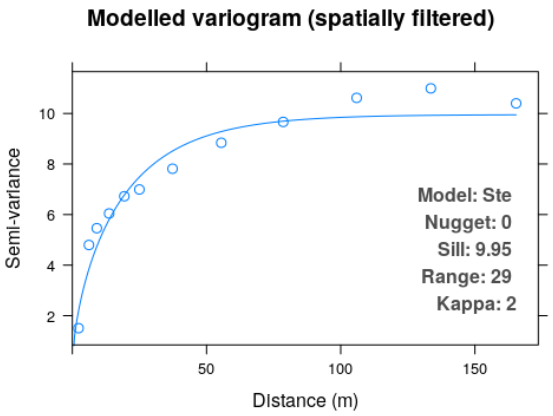
(a) 2014 un-filtered (nugget/sill = 0.18)



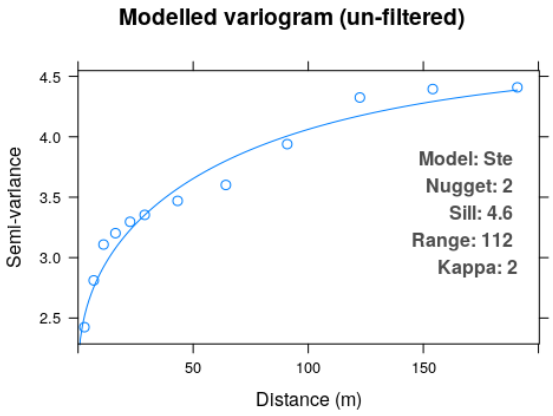
(b) 2014 filtered



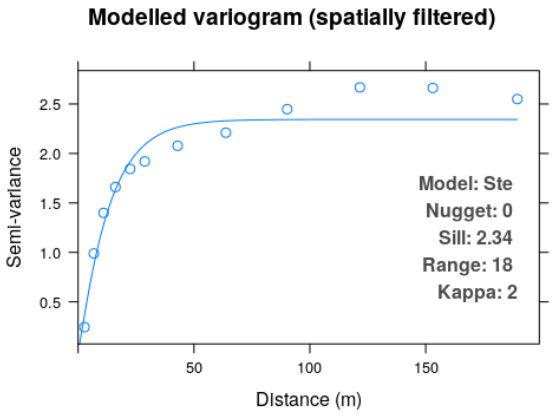
(c) 2015 un-filtered (nugget/sill = 0.91)



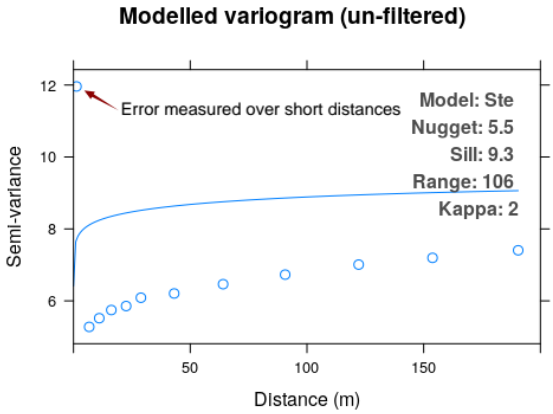
(d) 2015 filtered



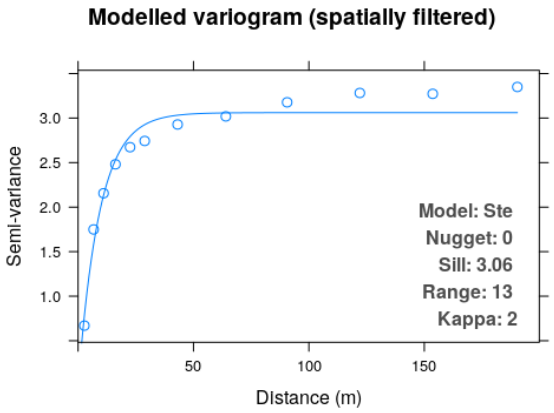
(e) 2017 un-filtered (nugget/sill = 0.43)



(f) 2017 filtered



(g) 2018 un-filtered (nugget/sill = 0.59)



(h) 2018 filtered

Figure 4-1 Comparisons of the experimental variograms for the individual-year yield monitor data before (a, c, e, g) and after filtering (b, d, f, h) using the customised spatial filtering algorithms.

To evaluate how spatial filtering influenced the accuracy of kriging, the interpolated yield values were compared with the observed yield values, and the RMSEs were obtained. Table 4-2 showed that the spatial filtering method that was applied reduced the RMSEs for all available years (2014, 2015, 2017, and 2018), particularly for the 2014 and 2015 yield data, in which the RMSEs were reduced substantially from (4.41 – 4.59) to (0.96 – 1.43); and from (6.61 – 7.27) to (1.34 – 1.43), respectively, suggesting that the filtering algorithm was effective at improving the kriging prediction for mapping spatial yield. The RMSEs generated from fitting different variogram models (Table 4-2) were similar, but the Matern model "Ste" generally produced smaller errors than the others.

Table 4-2 10-fold cross-validation root mean squared errors (RMSEs) of ordinary kriging (fitted with different variogram models fitted: Spherical "Sph"; Exponential "Exp"; Gaussian "Gau"; Matern "Ste") obtained by comparing the interpolated values and the observed values

Year	10-fold cross-validation RMSE of ordinary kriging							
	Before filtering				After filtering			
Fitted variogram models	Sph	Exp	Gau	Ste	Sph	Exp	Gau	Ste
2014	4.41	4.43	4.59	4.58	1.03	0.96	1.43	0.98
2015	7.27	7.13	7.27	6.61	1.43	1.43	1.34	1.43
2017	1.50	1.42	1.41	1.37	0.36	0.36	0.44	0.35
2018	2.07	2.04	2.13	2.05	0.69	0.64	0.72	0.64

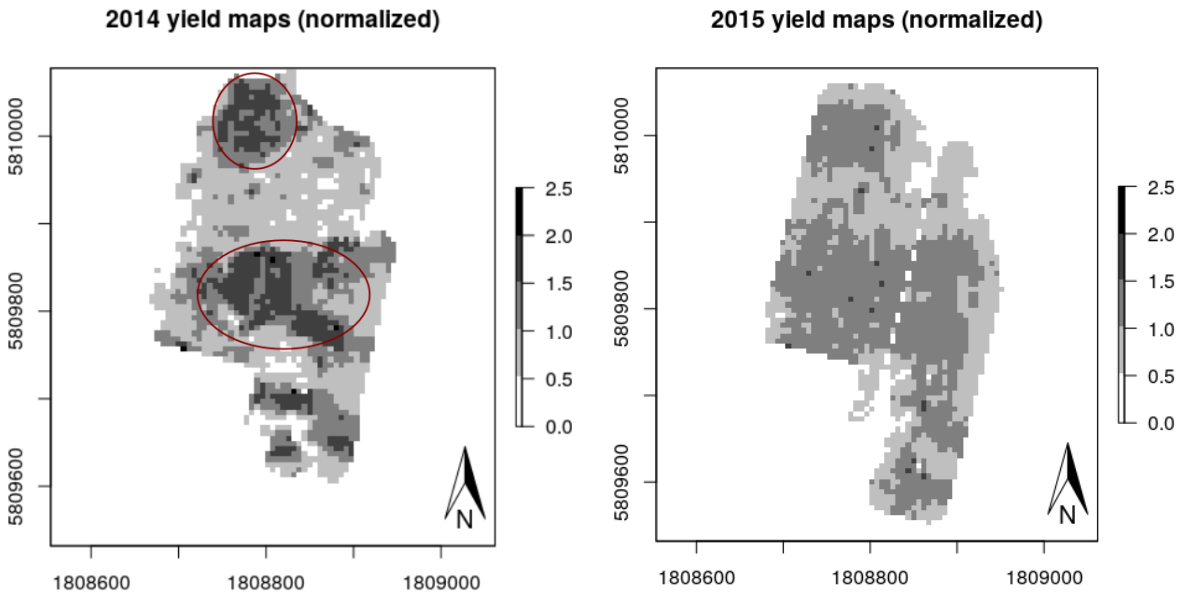
To improve the precision of subsequent analysis, the filtered yield data were used in yield mapping (section 4.2), delineating static zones (section 4.3), and as the response variable in the supervised machine learning modelling analysis (section 4.4).

4.3 Examining spatial yield and soil variability

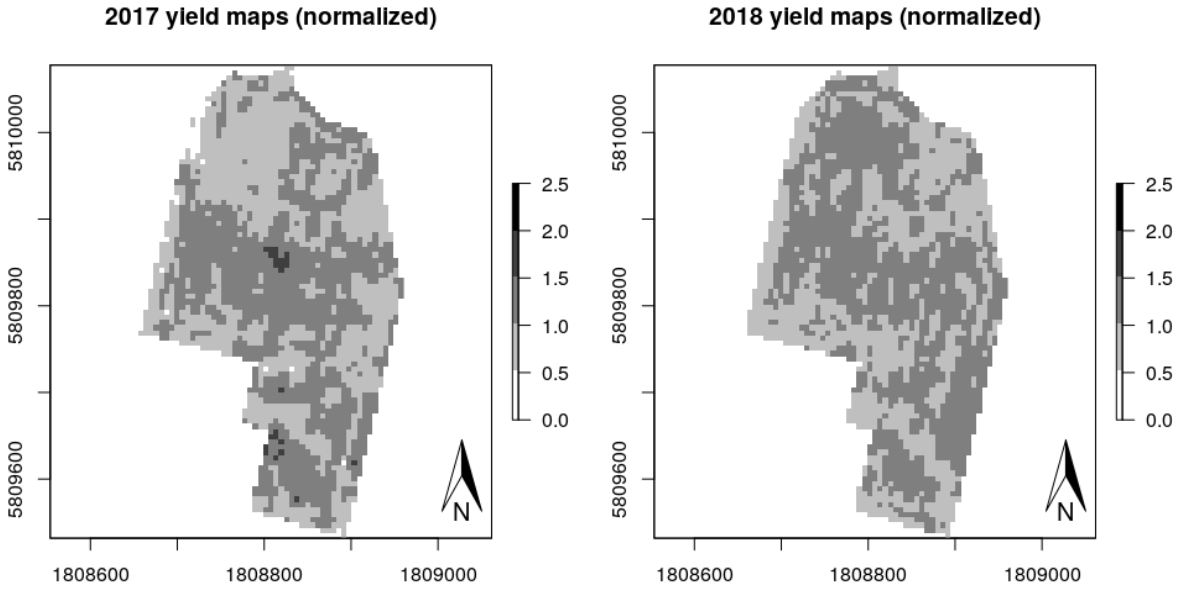
4.3.1 Mapping spatial data

4.3.1.1 Historical yield maps (Response variable)

To align the values (yield monitor data, soil EC, OM and elevation) originally sampled at different locations within-field, the spatial yield data points for the individual years (2014, 2015, 2017, and 2018) were interpolated into maps for geospatial analysis. Compared to the yield maps of 2015, 2017, and 2018, the 2014 yield map produced a strong "cluster" pattern as shown in Figure 4-2a. These clusters can be seen marked by the red circles. The yield within the "clusters" ranged from 1.5 – 2 times the field average (10 - 16 t/ha), whereas in other areas it mostly ranged from 0.5 - 1 times the field average (4 - 7 t/ha). To attempt to explain the yield variability, the rainfall data for the 2013/14 crop growing period was presented (Figure 4-3).



(a) (b)



(c) (d)

Figure 4-2 Historical yield maps (normalised in relative to the field average yield for the year): (a) 2014 yield map (the red circles represent two observed clusters); (b) 2015 yield map; (c) 2017 yield map; (d) 2018 yield map. (Projection: NZTM in meters)

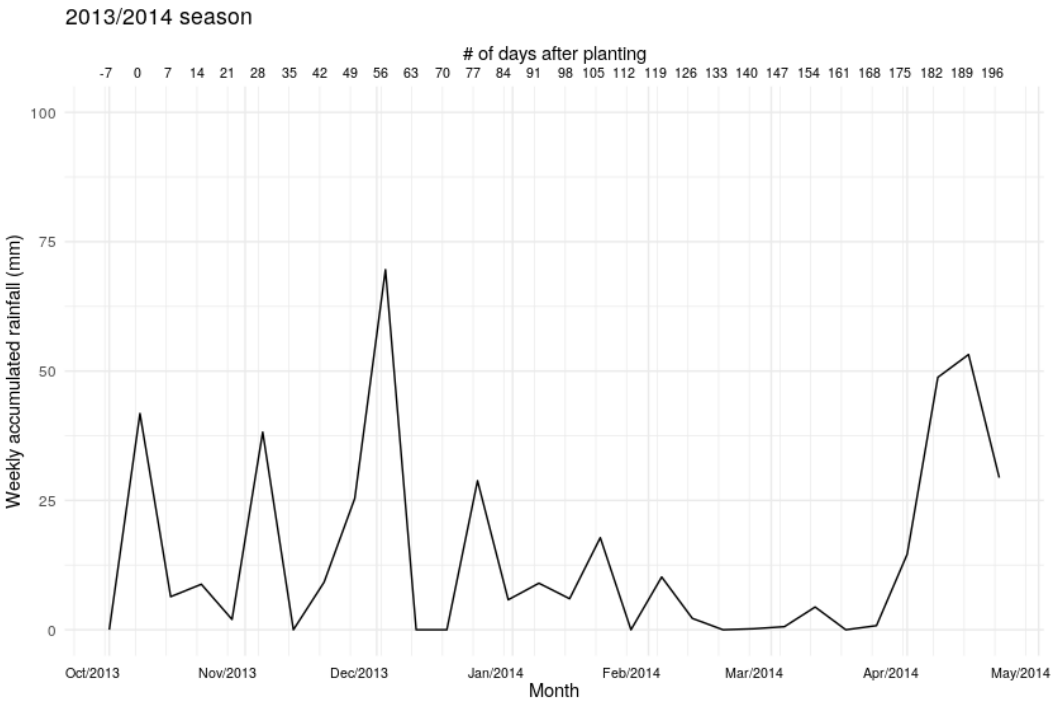


Figure 4-3 Weekly total rainfall for the 2014/2015 season (the crop was planted on October 8, 2013, and harvested on May 12, 2014)

The descriptive statistics for each yield map are summarised in Table 4-3. The CVs confirm the magnitude of the yield variability within-field for each year. The 2014 yield map produced the largest CV of 0.42, indicating the largest yield variability within-field for that year. Other years produced smaller CVs ranged from 0.16 - 0.26, i.e. less variability. A large CV of 0.33 was also produced for the combined yield map from all available years. Given that this CV is larger than those produced in 2015, 2017, and 2018, the yield map of 2014 (with greater variability and a larger CV), had a greater weighting in overall yield variability. This result suggests that in addition to overall spatial yield variability, there was also a large temporal yield variability to be considered when planning effective crop management. This also suggests that temporal factors such as seasonal rainfall should be incorporated into the development of modelling yield variability.

Table 4-3 Statistics of interpolated yield values from yield monitor data (mean, standard deviation [sd], coefficient of variation [CV], minimum [min] and maximum value [max])

Year	min	max	mean	sd	CV
			t/ha		
2014	1.59	17.37	7.79	3.25	0.42
2015	3.69	22.13	12.42	3.19	0.26
2017	2.85	12.49	7.82	1.63	0.21
2018	5.99	16.08	11.08	1.74	0.16
All year	1.59	22.13	9.73	3.21	0.33

To determine the overall spatial yield variability, the yield maps of four years (2014, 2015, 2017 and 2018) were averaged. Figure 4-4a showed that the variation of the average yield ranged from 0 t/ha to 16 t/ha. This large yield variability challenges the conventionally uniform treatment of the field as a single unit, and suggests the potential of optimising the application of crop management inputs such as seed and fertiliser, based on within-field yield potential. There was an overall cluster pattern of yield, which can be explained by analysing soil variability (section 4.3.2.2).

The coefficients of variations (CVs) for the multiple-year yield maps are shown in Figure 4-4b. This map represents the rate of change in the yield pattern over the four years. The higher the CV value, the more temporally unstable the yield was over the years. The distribution of the CVs was skewed to the right, suggesting that the yielding pattern was mostly stable (Figure 4-5). Eighty per cent of the areas in Figure 4-4b had a CV between 10% and 40%.

To reveal the underlying spatial pattern from Figure 4-4b, CV threshold is adjusted. The CV threshold that divides the field into two equal sizes was approximately 18.5%. When this threshold was reduced by 5%, the field became mostly dominated by unstable yields and no clear spatial pattern was observed (Figure 4-4c). When the 18.5% threshold was increased to 23.5%, some noticeable unstable yielding pattern appeared (Figure 4-4d). The unstable yield area west of the green line (Figure 4-4d) was near a shelterbelt, suggesting crop shading and crop moisture stress induced by the shelter trees. The results from this CV analysis described in the previous paragraph suggests a more complex local environment contributing to spatiotemporal yield variability.

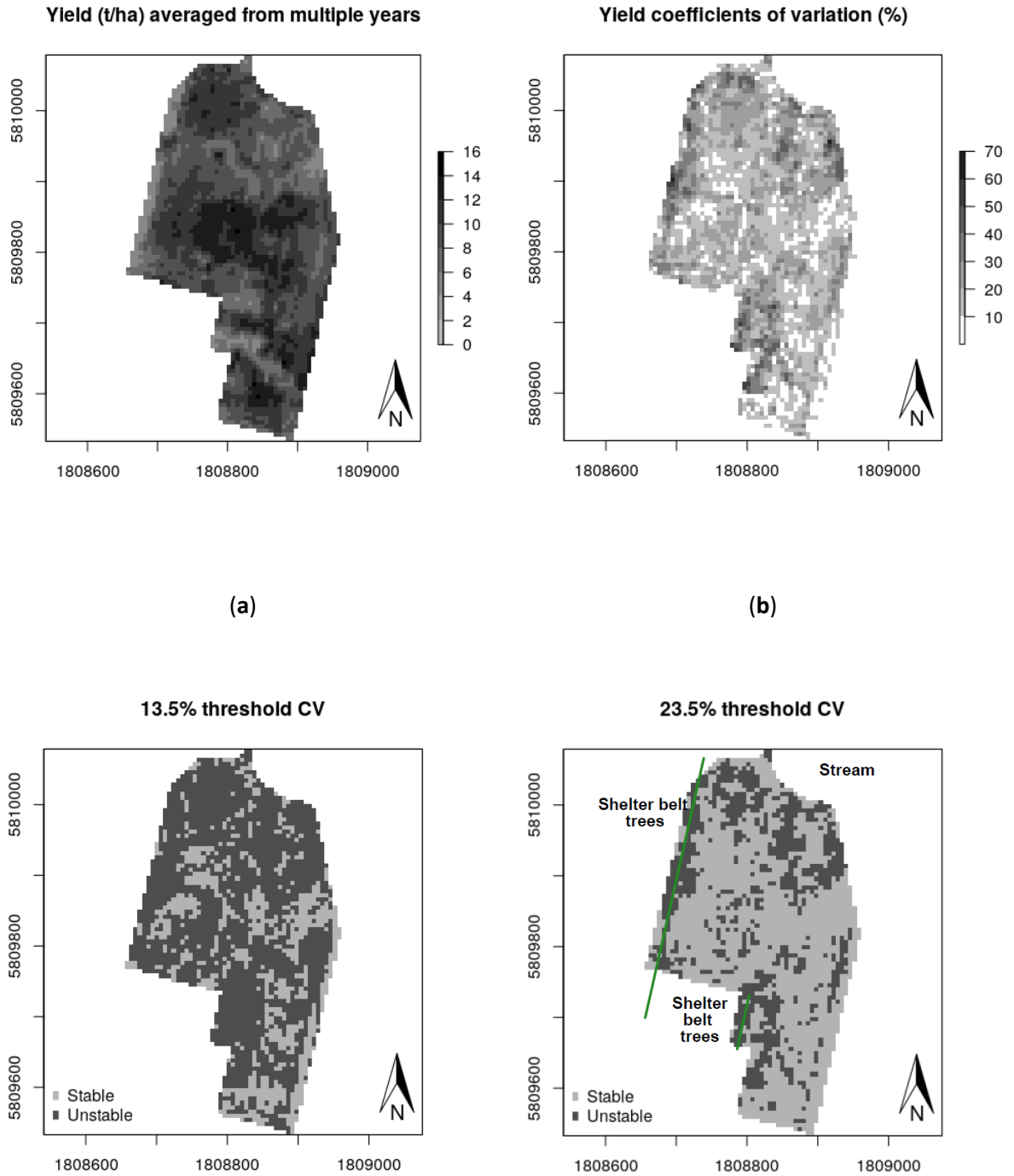


Figure 4-4 Combined maps from multiple-year maize yield maps (2014, 2015, 2017 and 2018): (a) yield average map (shows the overall yield productivity at each location); (b) Yield coefficients of variation (CV %) map (c) Yield CV map with unstable and stable yielding zones divided by 13.5% CV threshold (d) Yield CV map with unstable and stable yielding zones divided by 23.5% CV threshold (The left of the green line represents the hedgerow trees).

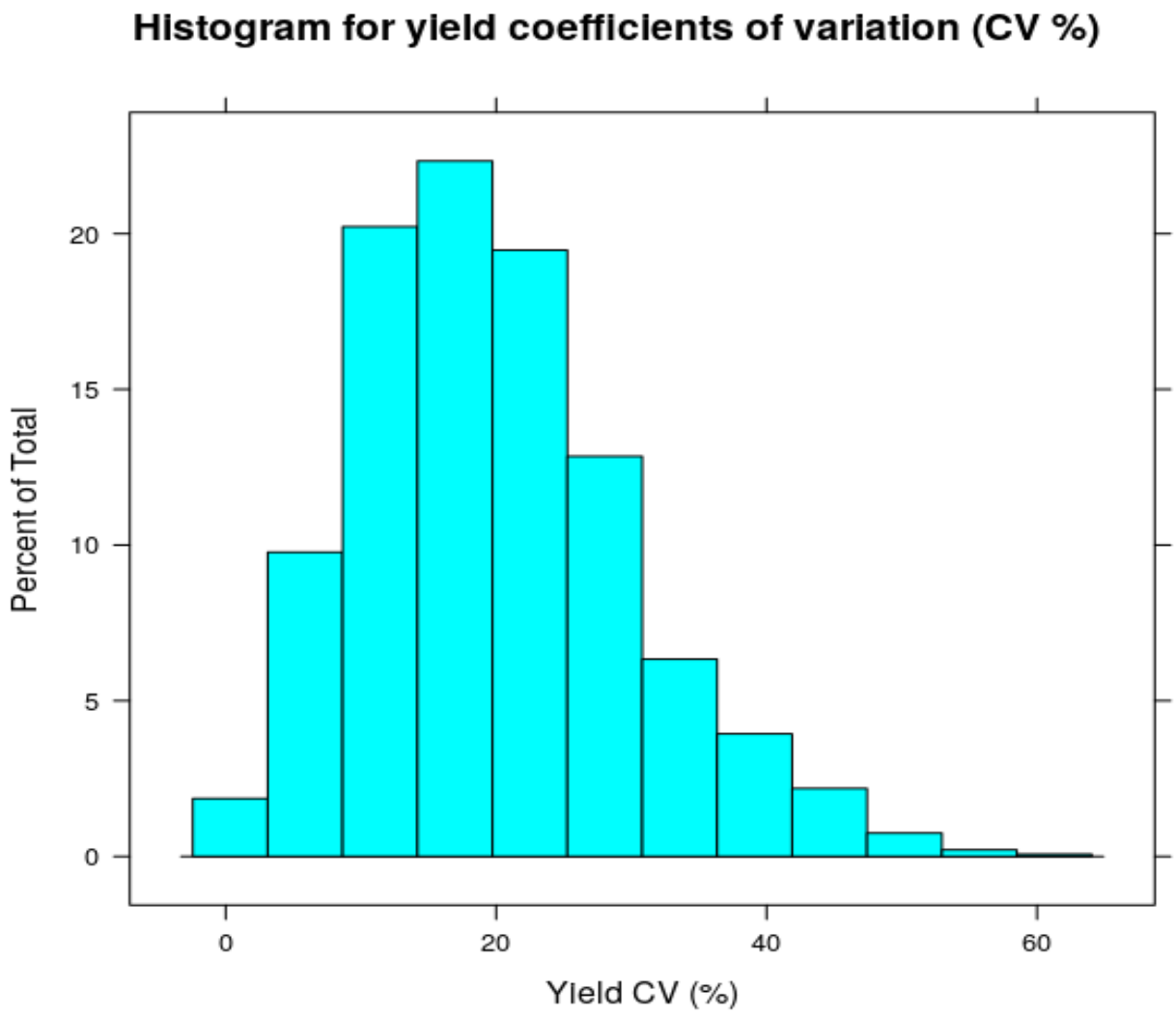


Figure 4-5 The histogram of the yield CV distribution.

4.3.1.2 Soil EC, organic matter, elevation maps (predictors)

To determine the cause of spatial yield and to compare the predictors (soil EC, OM, and elevation), data points were interpolated into maps (Figure 4-6) using ordinary kriging. To evaluate spatial soil variability, the parameters of the modelled soil variograms are presented (Table 4-4). The nugget variances for these variograms are all close to 0, suggesting few measurement errors. Low RMSEs were produced relative to the mean value of the data, indicating good interpolation accuracies.

Table 4-4 Parameters of the modelled variograms (models, nugget, sill, range and 10-fold cross-validation RMSE of spatial interpolation) for the spatial data

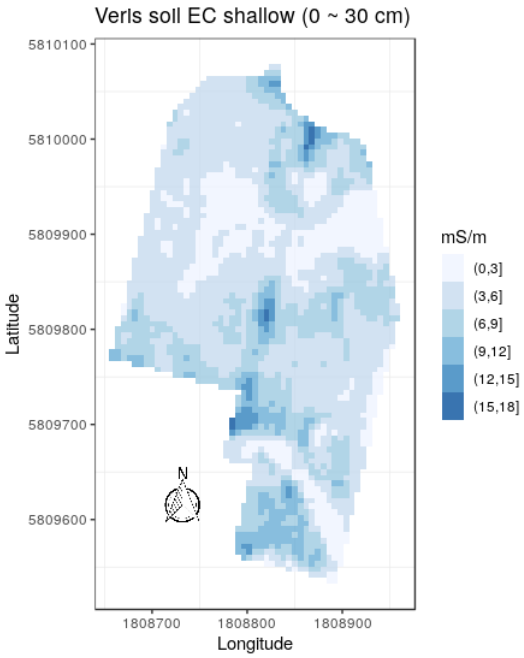
	Fitted variogram models	Nugget	Sill	Range m	RMSE	Mean	sd	N
EC shallow	Ste	0.00	6.89	38	0.25	5.4	2.6	4346
EC deep	Ste	0.01	19.22	41	0.75	6.2	4.3	4346
Elevation	Ste	0.01	206.59	137,244	0.11	54.7	0.8	4854
Soil organic matter	Ste	0.00	0.00	23	0.02	2.5	0.1	4346

Soil EC shallow (0-30 cm) and soil EC deep (0-90 cm) were strongly correlated (r = 0.89) (Figure 4-6a, b). Soil EC shallow values ranged from 1.2 – 16.3 mS/m, suggesting a soil texture composition of mainly sand and silt. Soil EC deep values ranged from -8 to 32.2 mS/m, suggesting larger texture ranges in the deeper soil profile (Grisso et al., 2005). Both soil EC maps showed a clear spatial "banding" pattern with values ranging from 0 to 3 mS/m, indicating coarse-textured soils. To correlate EC with soil texture, soil core sampling was undertaken at different depths with subsequent soil particle size analysis undertaken (see section 3.3.5).

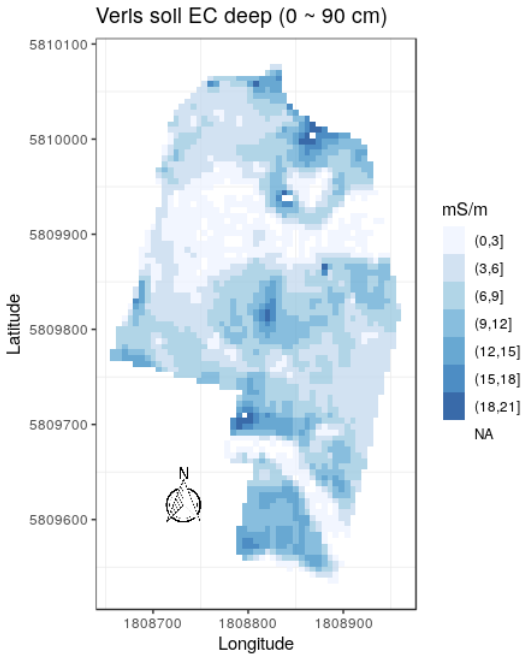
An elevation map (Figure 4-6c) shows that there is a gradual decrease of elevation by approximately 4 m from the South-eastern boundary northward over the 400 meters at the adjacent stream on the northern boundary.

The soil organic matter map (Figure 4-6d) shows a little variation (2.4 - 2.65%), but there are some noticeable "strips", aligned in the direction in which farm operations (such as planting) occurs. These "strips" may be an artefact of the historical strip-till crop establishment in the field, which only cultivates narrow strips for seed placement while leaving most surface residues on the surface.

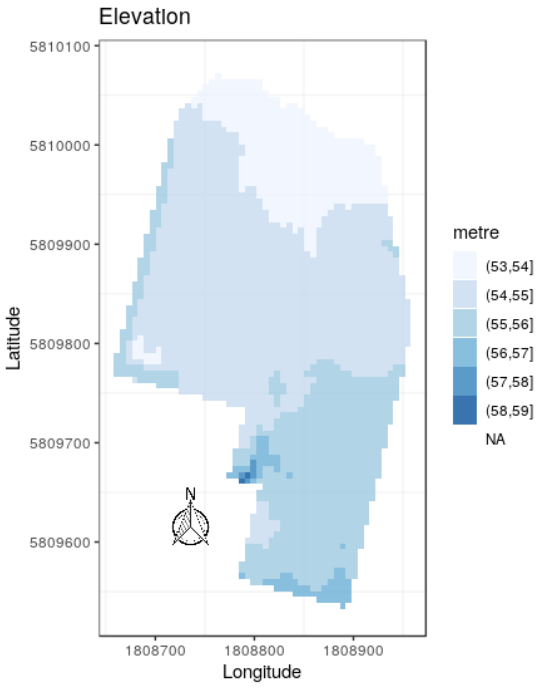
These maps (Figure 4-6) were used as yield predictors in multivariate modelling analysis (see section 4.4).



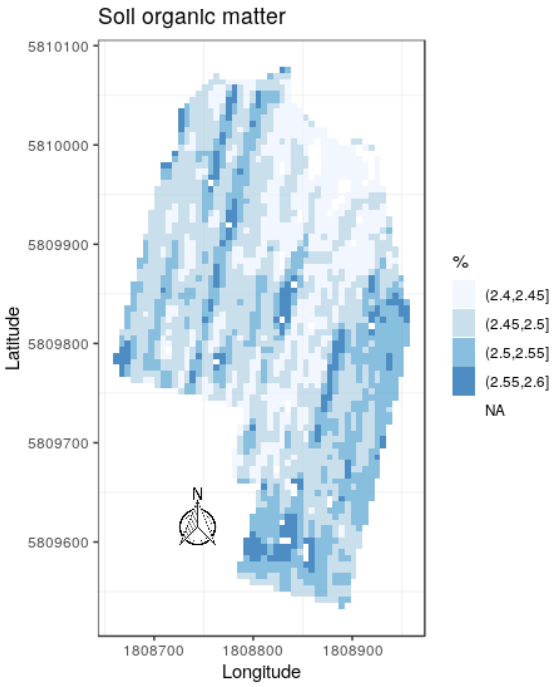
(a) Soil EC shallow



(b) Soil EC deep



(c) Elevation



(d) soil organic matter

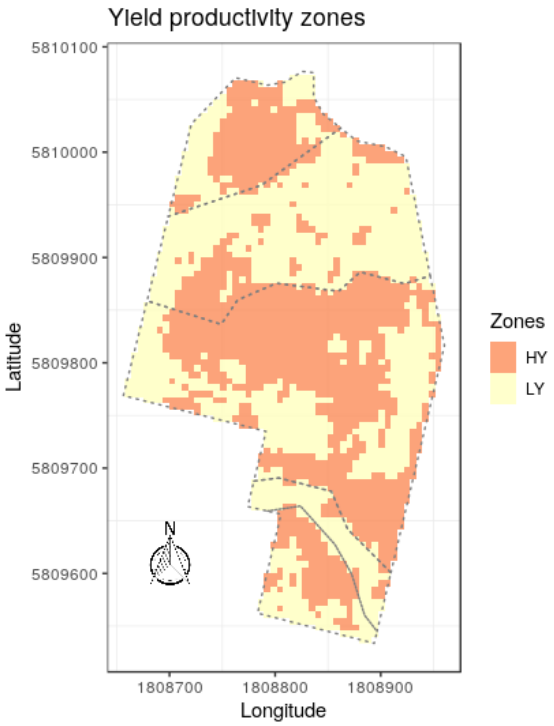
Figure 4-6 Potential yield predictors (soil EC & OM) derived from Veris MSP-3 soil sensor and RTK GPS (Projection: NZTM in meters)

4.3.2 Potential crop management zones (static)

4.3.2.1 Yield productivity zones (YPZ)

To delineate yield MZs, historical yield data (2014, 2015, 2017, and 2018) was classified into relatively high-yielding [HY] and low-yielding [LY] zones based on the average yield over the four years. These yield MZs could assist in the application of different management inputs (see Method section 3.4.3.1). Figure 4-7b shows that the LY zones are visually consistent with the observed "banding" visible from a Google Earth image that illustrates differential growth patterns (Figure 4-7a). This visual consistency demonstrates that the spatial yield variability is consistent with aerial imagery.

(a)

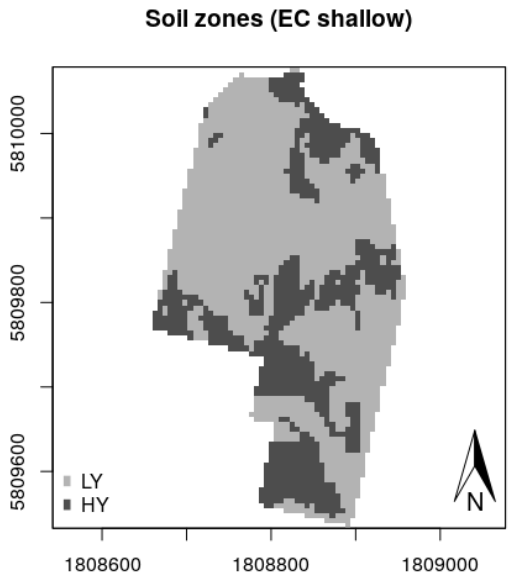


(b)

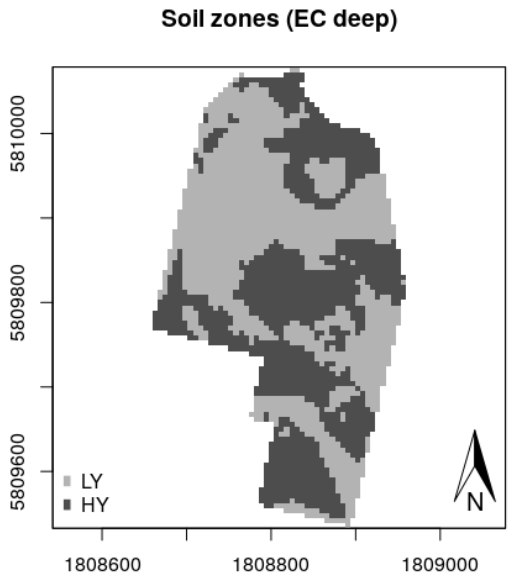
Figure 4-7 (a) the observed "banding" pattern of the crop (delineated by the dotted lines) on a Google Earth image (11 March 2016) and (b) relative yield productivity zones (relatively high yielding [HY] potential and relatively low yielding potential [LY] based on historical average yield) (Projection: NZTM in meters).

4.3.2.2 Soil zones (SZ)

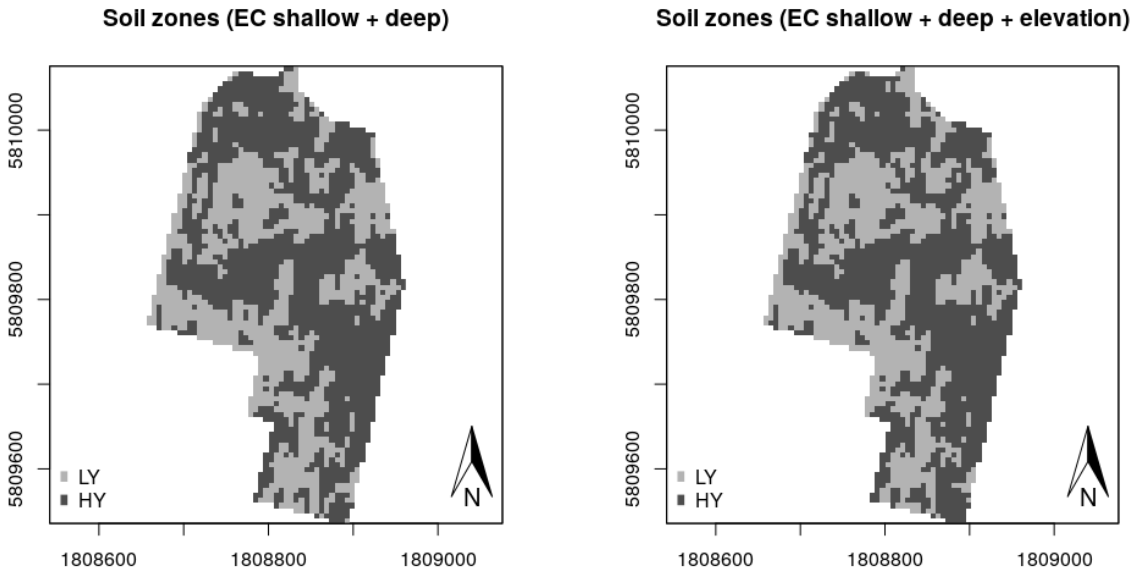
To determine the best predictors for spatial yield, all statistical combinations of soil EC (deep and shallow) and elevation maps were delineated into soil zones. This enabled comparison between both the yield zones and soil zones to examine correlation (Figure 4-8). HY and LY zones were labelled based on the average yield for that zone over the four years. There was a distinct visual consistency of pattern between all four maps and this is quantified in section 4.3.3. Adding elevation to clustering of the soil EC maps did not appear to affect the spatial yield zone pattern (Figure 4-8d). The statistical comparison between soil zones and yield zones are presented in Table 4-5.



(a)



(b)



(c) (d)

Figure 4-8 Soil zones delineated from the combinations of soil EC and elevation maps using fuzzy c-means clustering (Zones are labelled as relatively high yielding [HY] potential and relatively low yielding potential [LY] based on historical average yield) (Projection: NZTM in meters)

4.3.2.3 Crop reflectance zones (CRZ)

To enable comparison of the delineated zones, Sentinel-2 multispectral-band images for the field were downscaled (from 10 m to 6 m of the interpolated yield maps) using ordinary kriging. The delineated crop reflectance zones (CRZ) are presented in Figure 4-9. To compare these with the yield zones, the clusters derived from the multispectral reflectance data were labelled HY and LY based on the historical yield average for each zone. This zone pattern was visually similar to those observed on a Google Earth image (delineated as a dotted line in Figure 4-9). This suggests that crop reflectance zones can be used as an indicator of field variability.

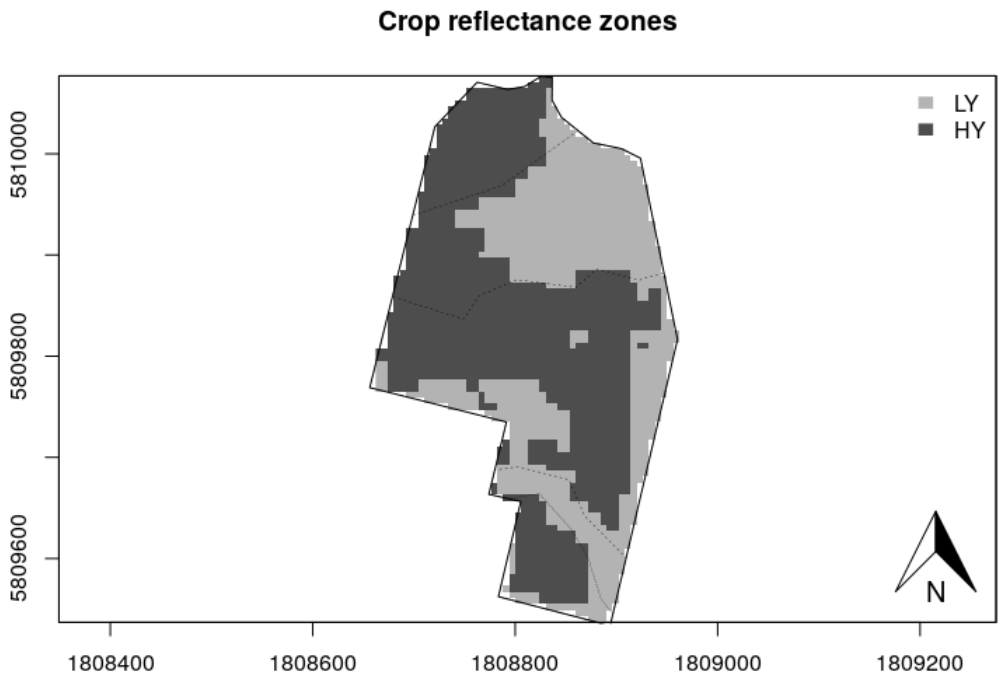


Figure 4-9 Crop reflectance zones delineated from multi-date satellite images (relatively high yielding [HY] potential and relatively low yielding potential [LY] labelled based on historical yield average) and the visual "banding" pattern (delineated by the dotted lines) (Projection: NZTM in meters)

4.3.3 Statistical comparison of zone pattern

To statistically compare the zone patterns, kappa coefficients were calculated for each pairwise combination of the zone maps. Table 4-5 shows that some degree of similarity between the static zones (yield productivity zones, soil zones and crop reflectance zones) was found by comparing their spatial patterns. The best area agreement was found between crop reflectance zones and yield productivity zones (71%), which suggests that there is a potential to use Sentinel-2 reflectance images as reliable proxies for yield maps for delineating yield MZs when yield data are not available. The second-best agreement was found between soil zones (including elevation) and yield productivity zones (63%), which suggests that the soil EC and elevation have some contribution to yield variability.

To determine the statistical confidence of that spatial alignment, Kappa coefficients were calculated (see Method 3.4.3.4). Kappa coefficients were generally low (0.04-0.42), which suggests that agreement between yield productivity zones and any soil zone maps may have occurred by chance.

Table 4-5 Areal agreements (%) between zones. Corresponding kappa coefficients are in brackets (0–0.20 not reliable; 0.21–0.39 minimal; 0.40–0.59 weak, 0.60-0.79 moderate, 0.80-0.90 strong, above 0.90 almost perfect alignments (McHugh, 2012)).

Zones	SZ (elevation incl.)	SZ (EC shallow)	SZ (EC deep)	SZ (EC shallow + deep)	CRZ
YPZ	63% (0.27)	53% (0.04)	58% (0.16)	63% (0.26)	71% (0.42)

To determine if the delineated zones (yield productivity zones, soil zones and crop reflectance zones) are able to represent different yield levels, historical yield values were averaged for each zone and compared using ANOVA tests. The zones derived from the different spatial datasets (soil EC/elevation, reflectance, and yield) were most effective at separating the yield levels over the four years, as indicated by the *p*-values ($p < 0.05$) (Table 4-6).

However, there were levels of variation in the average yield between HY and LY. For crop reflectance zones, the variation in the average yield between zones was greater for 2014, 2015 and 2018 compared to 2017, which could be related to relatively good soil moisture status and more plant-available water in all other years than 2017.

For soil zones, there was no significant difference in the average yield between the soil zones for 2014 and 2015. This statistical significance analysis reaffirms the temporal effect of EC on spatial yield and the need to incorporate weather data into yield zone prediction models.

Table 4-6 Average yields in static zones for each year and the p-value derived from ANOVA tests (subsampled N = 100).

Zones	SZ (elevation incl.)			YPZ			CRZ		
	HY	LY	*p-value*	HY	LY	*p-value*	HY	LY	*p-value*
year									
2014	9.6	7.0	<0.001	10.9	5.5	<0.001	9.9	6.2	<0.001
2015	14.2	12.6	<0.05	15.0	11.2	<0.001	14.5	11.2	<0.001
2017	8.1	7.7	ns	8.7	7.1	<0.001	8.2	7.5	<0.05
2018	11.5	10.8	ns	12.1	10.4	<0.001	11.5	10.8	<0.01
Area (ha)	5.8	4.3		4.9	5.1		5.9	4.1	
Average (t/ha)	10.5	9.4		11.7	8.6		11.0	8.9	
Total yield (t)	60.9	40.4		57.1	44.2		63.0	36.4	

4.3.4 Results of soil particle size analysis

4.3.4.1 Soil particle size fractions at various depths

To verify the soil EC relationship with soil texture, soil sample cores were taken from each yield productivity zone and analysed for particle size composition. Table 4-7 shows that particle size fractions varied both spatially and at depth. Spatially, the average sand fraction in the top 30 cm ranged from 24 - 64% for the six locations; the average silt fractions ranged from 31-69%; and the average clay fractions ranged from 12-24%. At depth, there were large variations, particularly for the silt and clay fractions at HY1 and at LY1 (Table 4-7; sampling locations refer to Figure 3-4), as indicated by the coefficients of variation (CV) for HY1 (silt 30%, clay 85%); LY1 (silt 27%, clay 44%). However, no obvious pattern can be identified associated with these variations at depth.

HY3 (relatively high yielding) had the highest sand content for all depths (57-74%) compared to the other sampling locations. LY2 (relatively low yielding) had the highest silt (72-75%) and clay (28-31%) content at 20-30 cm depth. LY3 had the highest silt (69-72%) and clay (22-28%) at 5-20 cm depth. The results for LY2 and LY3 could suggest potentially poorly drained soils with fine texture.

Table 4-7 Particle size distribution (sand, silt and clay) at various depths (5-10 cm, 10-15 cm, 15-20 cm, 20-25 cm and 25-30 cm)

Core label	HY1	HY2	HY3	LY1	LY2	LY3
Sand fraction						
5-10 cm	0.23	0.36	0.57	0.55	0.25	0.25
10-15 cm	0.27	0.30	0.60	0.57	0.25	0.23
15-20 cm	0.18	0.31	0.74	0.54	0.25	0.26
20-25 cm	0.25	0.25	0.62	0.48	0.22	0.36
25-30 cm	0.24	0.27	0.64	0.40	0.20	0.28
profile average	0.24	0.30	0.64	0.51	0.23	0.28
CV †	0.14	0.14	0.1	0.14	0.1	0.18
Silt fraction						
5-10 cm	0.34	0.47	0.32	0.26	0.66	0.71
10-15 cm	0.63	0.61	0.32	0.37	0.66	0.72
15-20 cm	0.39	0.60	0.33	0.33	0.68	0.69
20-25 cm	0.67	0.66	0.30	0.47	0.72	0.55
25-30 cm	0.67	0.66	0.30	0.52	0.75	0.68
profile average	0.54	0.60	0.31	0.39	0.69	0.67
CV	0.3	0.13	0.04	0.27	0.06	0.1
Clay fraction						
5-10 cm	0.06	0.15	0.12	0.16	0.25	0.26
10-15 cm	0.00	0.15	0.13	0.15	0.26	0.22
15-20 cm	0.12	0.20	0.13	0.08	0.27	0.28
20-25 cm	0.27	0.18	0.13	0.18	0.28	0.24
25-30 cm	0.27	0.25	0.11	0.29	0.31	0.23
profile average	0.14	0.18	0.12	0.17	0.27	0.24
CV	0.85	0.22	0.07	0.44	0.08	0.1

† Coefficient of variation

To determine the precision of analyses using the pipette method (Claydon, 1989), replicate analyses were conducted using the soil core subsamples. The average coefficients of variation (CV) calculated ranged from 0.02 to 0.04 for sand, from 0.07 to 0.15 for silt and from 0.07 to 0.32 for clay. These low CV's indicated that the results from each measurement were consistent.

4.3.4.2 Soil particle size distribution influence on soil EC

To examine if soil EC data can be used as a proxy for soil texture, soil EC was correlated to different particle size fractions using Pearson's correlation coefficient (rho). The soil EC shallow data (0 - 30 cm) were strongly positively correlated to the clay fraction (<2 microns) at 5-10 cm depth (r = 0.94), followed by the clay fraction at 10-15 cm depth (r = 0.82) (Figure 4-10).

However, the correlations decreased at the 15 – 30 cm soil profile depth (r = 0.16 – 0.68). A positive correlation was also found between the soil EC shallow (0-30 cm) and the silt fraction at 5-10 cm depth (r = 0.75). There were mostly negative and weak correlations between soil EC shallow and the sand fractions. Therefore, the within-field areas with low soil EC are likely to be associated with relatively coarse-textured soils and thus lower water holding capacity (Figure 4-6a, b).

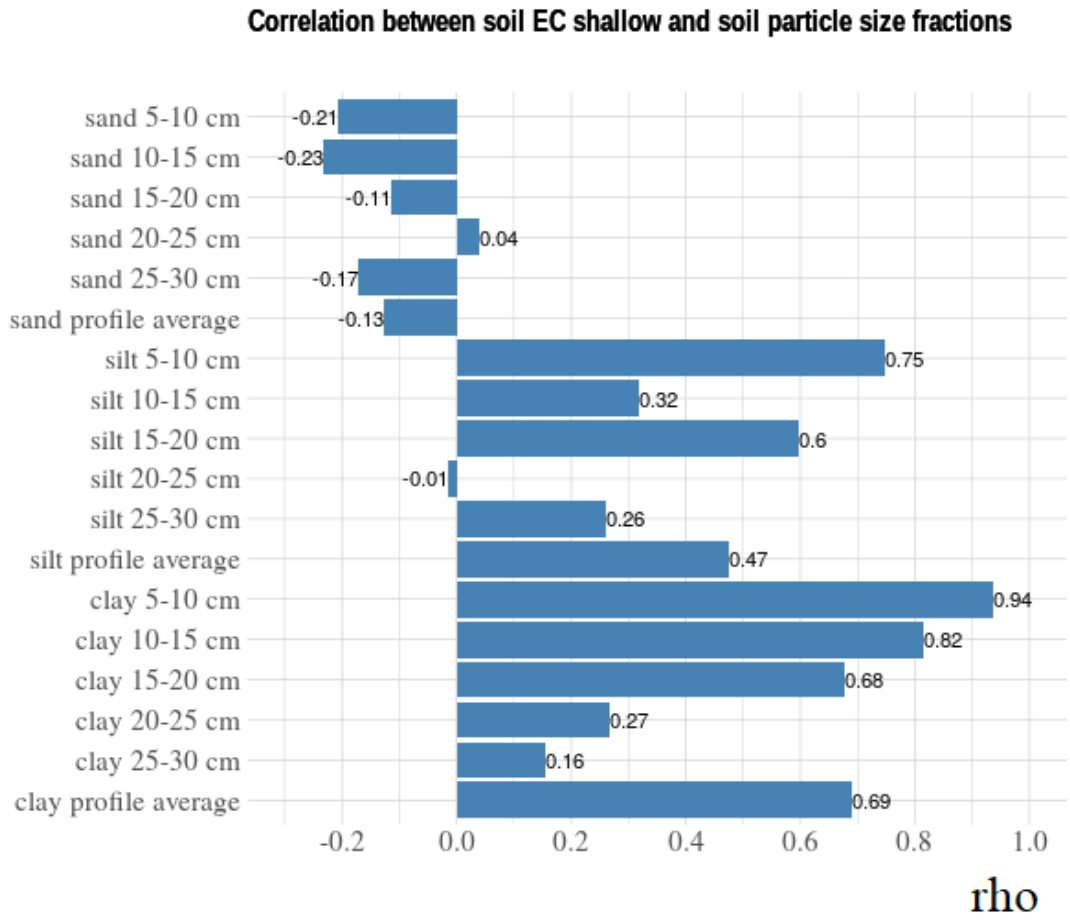


Figure 4-10 Correlation between soil EC shallow (0-30 cm) and soil particle size fractions (sand, silt and clay) at depths (5-10 cm, 10-15 cm, 15-20 cm, 20-25 cm, 25-30 cm).

4.3.4.3 Soil particle size distribution influence on crop yield

To identify multivariate relationships between soil texture and crop yield, PCA was applied to the dataset containing multiple-year yield data (2014, 2015, 2017, and 2018), soil EC (shallow and deep), and soil particle size fractions (sand, silt, and clay) as investigated in section 4.3.4.1.

The first two major principal components (PCs) explained 79% of the total variation. LY2 and LY3 (relatively low yielding locations shown in Figure 3-4, determined based on Figure 4-7) showed similar properties in terms of their soil particle size fractions. LY2 and LY3 were dominated by silt and clay in the surface soil (5-30 cm), as was EC shallow (Figure 4-11).

In PC1, the multiple-year yield (2014, 2015, 2017, and 2018) were negatively correlated to the EC data (both shallow and deep), suggesting that higher EC values (and potentially saturated soils) may have contributed to low yield productivity at LY2 and LY3 (Figure 4-11).

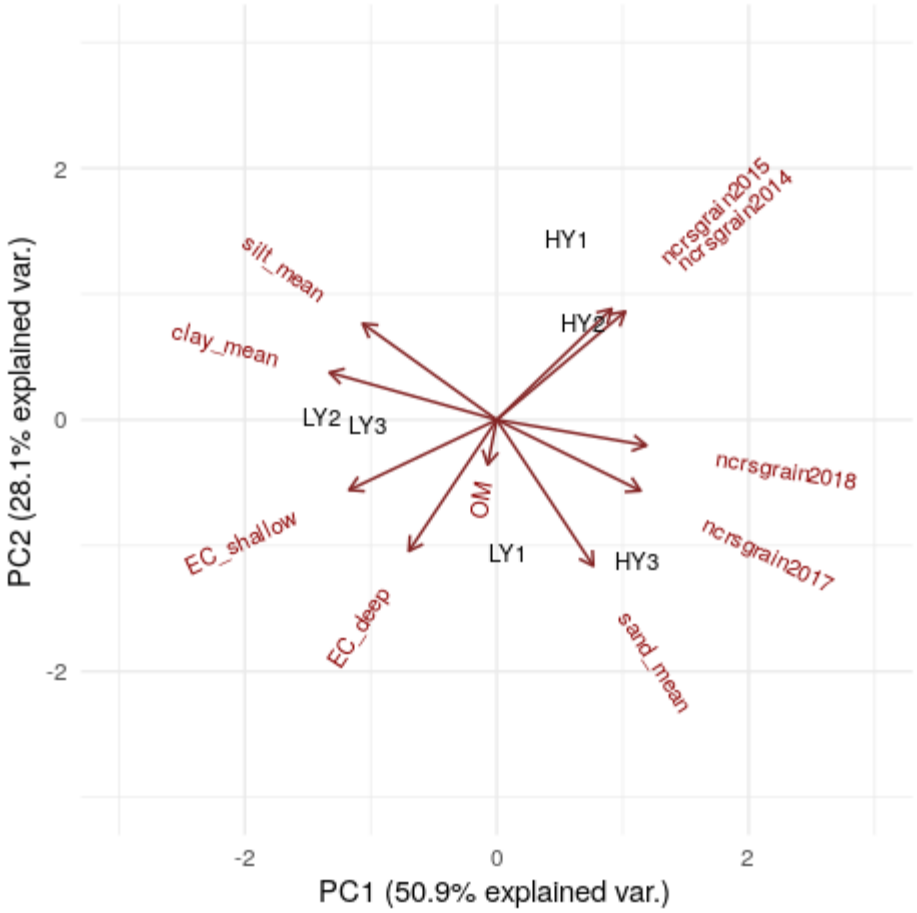


Figure 4-11 PCA biplot shows the clusters of samples (HY, LY) based on their similarity (Each PC is the linear combination of the original variables: profile average sand, silt, clay, OM, EC shallow, EC deep, multiyear yield data 2014 – 15 and 2017 – 2018). The longer the arrow of a variable, the greater contribution that variable has in this two-dimensional (PC1 + PC2) space.

To determine the effect of soil EC on yield for each year, quadratic models were fitted (Kitchen et al., 2003). The results (provided in Table 4-8) shows that the quadratic model provided a good fit for 2015 and 2017 (R^2 = 0.76 and R^2 = 0.92, respectively), suggesting that yield increases as soil EC shallow increases. High soil EC responses are related to the higher water-holding capacities of finer-textured soils. Therefore, plants located in areas with higher soil EC are less likely to encounter moisture stress. However, at this site, as EC increased, yield decreased, possibly due to impeded drainage and/or a perched water table. The quadratic models produced a poorer fit for 2014 and 2018 (R^2 = 0.44 and R^2 = 0.56, respectively), suggesting that other factors such as fertiliser application variability could have played an important role in the yield response to soil EC shallow.

Table 4-8 Equations for the average yield response to soil EC shallow and R^2

Year	Equation	**R^2**
2014	$y=4.93+0.915x-0.0424x^2$	0.44
2015	$y=8.03+1.92x-0.136x^2$	0.76
2017	$y=6.55+0.363x-0.0147x^2$	0.92
2018	$y=9.67+0.442x-0.0272x^2$	0.56

4.4 Results of multivariate modelling

4.4.1 Prediction performance

4.4.1.1 Multiple year analysis

To evaluate yield prediction performance and to compare different models, the "split-sample" approach (75:25 data partitioning) was used (section 3.4.5.2). The nonlinear models (feedforward neural network [FFNN], classification and regression tree [CART], random forest [RF], XGBoost and Cubist) produced more accurate yield prediction (R^2 = 0.36 - 0.72) than the stepwise multiple linear regression (SMLR) model (R^2= 0.20 - 0.72), see Table 4-9. RF, XGBoost and cubist (R^2=0.44 - 0.72) produced relatively better prediction results than the SMLR and FFNN. This is to be expected as tree-based models are theoretically capable of handling data with multiple spatial and temporal interaction terms. The accuracy of CART was close to FFNN (R^2= 0.52 - 0.60), but CART required less computational power and required less training time than FFNN (both models are described in section 3.4.5).

To test this modelling method, more data from other fields are required. For most fields (NCRS, Field 3 and Field 5), the accuracy for training and validation (R^2) was close for all four nonlinear models, which suggested that the models were optimised (Hastie et al., 2009, pp. 228 - 230). For relatively smaller fields (Field 2 and Field 4), lower accuracies were observed for validation than training, suggesting that more fine-tuning of hyperparameters is required in 10-fold cross-validation.

For the pooled data, the best prediction of yield was provided by XGBoost (R^2 = 0.64 – 0.66). This result suggests that XGBoost could be more suitable for predicting yield with more variability.

Table 4-9.Prediction results (training and validation) provided by SMLR, FFNN, CART, RF, XGBoost and cubist in the multiple-year analysis (using data from the individual field).

	SMLR		FFNN		CART		RF		XGBoost		Cubist	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
NCRS (10 ha)												
Training	2.3	0.50	0.1*	0.53	2.1	0.55	2.0	0.60	2.0	0.63	2.0	0.61
Validation	2.3	0.51	2.2	0.52	2.0	0.60	1.9	0.63	2.0	0.62	2.1	0.57
Field 2 (5 ha)												
Training	2.5	0.46	0.1	0.49	2.5	0.48	2.5	0.48	2.4	0.49	2.5	0.48
Validation	1.9	0.46	1.9	0.46	2.2	0.36	1.9	0.46	2.1	0.44	2.0	0.45
Field 3 (14.5 ha)												
Training	2.9	0.45	0.1	0.60	2.8	0.48	2.3	0.64	2.3	0.66	2.5	0.59
Validation	2.9	0.36	2.5	0.52	2.6	0.50	2.1	0.64	2.1	0.67	2.5	0.52
Field 4 (7.8 ha)												
Training	1.1	0.82	0.05	0.82	1.2	0.81	1.1	0.81	1.1	0.82	1.1	0.82
Validation	1.3	0.72	1.3	0.72	1.3	0.72	1.3	0.72	1.3	0.72	1.3	0.71
Field 5 (24 ha)												
Training	2.1	0.16	0.1	0.36	1.8	0.36	1.7	0.44	1.7	0.46	1.8	0.39
Validation	2.0	0.20	1.7	0.39	1.7	0.39	1.5	0.50	1.6	0.50	1.6	0.45
Pooled												
Training	2.6	0.35	0.1	0.53	2.4	0.45	2.1	0.60	2.0	0.64	2.1	0.59
Validation	2.5	0.38	2.2	0.53	2.4	0.42	2.0	0.59	1.87	0.66	2.1	0.58

* the response variable in FFNN was scaled into 0-1

4.4.1.2 Leave-out-one-year analysis using individual-field data

Given the relatively small number of years, an internal cross-validation technique (i.e. withholding one year as test data while using all other years for training the model) was used to evaluate how these models predict unseen data. In this analysis, the yield prediction for the individual year (Table 4-10) was generally poor, producing low R^2 values. RF, XGBoost and Cubist produced better results than the other trained models (Table 4-10). RF produced the highest model fits (average R^2 = 0.08 - 0.50), followed by XGBoost (average R^2 = 0.06 - 0.39) and Cubist (average R^2 = 0.03 - 0.19).

Better prediction accuracies (average R^2 = 0.28 – 0.50) were produced for the larger fields (NCRS, Field 3 and Field 5) than Field 2 and Field 4 (average R^2 = 0.02 – 0.08). The highest prediction accuracies were produced by Field 3 for the years 2014, 2015, 2016, and 2017, using the RF models (average R^2 = 0.18 – 0.50). The prediction was poor for 2018 because Field 3 was planted with variable rate seeding in October 2017 and did not follow standard maize planting practice.

Table 4-10 Prediction results of SMLR, FFNN, CART, RF, XGBoost and Cubist in the leave-out-one-year analysis (using data from individual fields).

	SMLR	FFNN	CART	RF	XGBoost	Cubist	Observed yield		
				R^2					
Withhold							mean	CV%	Resampled *N*
NCRS (10 ha)									
2014	0.05	0.11	0.07	0.36	0.16	0.17	8.4	22%	103
2015	0.13	0.14	0.16	0.35	0.25	0.21	13.3	11%	103
2017	0.16	0.13	0.12	0.22	0.21	0.15	8.2	13%	103
2018	0.10	0.12	0.13	0.20	0.19	0.07	11.7	22%	103
Average	0.11	0.13	0.12	0.28	0.20	0.15			
Field 2 (5 ha)									
2014	0.02	0.02	NA	0.12	0.20	NA	16.7	22%	40
2015	0	0	NA	0.08	0.09	0.01	12.1	12%	40
2016	0.01	0	NA	0.06	0.01	0.15	13.6	10%	40
2017	0.05	0.09	NA	0.02	0.01	0.10	11.4	13%	40
2018	0.01	0.04	NA	0.01	0	0.03	10.7	30%	40
Average	0.02	0.03	NA	0.06	0.06	0.07			
Field 3 (14.5 ha)									
2014	0.46	0.45	0.09	0.44	0.15	0.32	9.5	62%	111
2015	0.49	0.42	0.36	0.69	0.59	0.46	7.3	49%	104
2016	0.57	0.50	0.50	0.74	0.63	0.01	9.1	29%	111
2017	0.47	0.40	0.39	0.54	0.51	0.08	9.5	28%	111
2018	0.03	0.03	0.03	0.09	0.09	0.02	11.1	14%	111
Average	0.40	0.36	0.27	0.50	0.39	0.18			
Field 4 (7.8 ha)									
2005	0	0	NA	0.11	0.06	0.05	16.6	11%	75
2007	0	0	NA	0	0	0.03	9.5	14%	75
2009	0.07	0.08	NA	0.06	0.03	0	10.2	10%	75
2010	0.04	0.01	NA	0.02	0.04	0.01	13.0	8%	75
2013	0.08	0	NA	0.27	0.19	0.02	10.5	10%	75
2015	0.03	0.09	NA	0.08	0.05	0.04	13.0	7%	71
2017	0.02	0.01	NA	0.04	0.06	0.03	10.8	7%	75
Average	0.03	0.03	NA	0.08	0.06	0.03			
Field 5 (24 ha)									
2008	0.05	0.01	0.18	0.49	0.43	0.37	13.8	22%	286
2009	0.08	0.02	0.37	0.54	0.48	0.13	12.7	11%	286
2010	0	0	0.06	0.08	0.10	0.06	11.7	13%	286
Average	0.04	0.01	0.20	0.37	0.34	0.19			

4.4.1.3 Leave-out-one-year analysis using pooled data

To determine the feasibility of predicting yield for an additional year for a field that has limited data based on models trained from pooled data, and if there is any improvement on the predictions based on models trained from the individual field, the predictions were evaluated by leaving out one set of data each term and using the remaining pooled data for training the models.

The results (Table 4-11) showed that the models (random forest and XGBoost) built using the pooled data generally produced poorer predictions than the models built using data collected from individual fields. Except for Field 3, the XGBoost models improved the R^2 of yield predictions of individual years by 0.01 to 0.23 using the pooled data. However, the results of XGBoost were mostly poorer compared with that of RF using data from individual fields. This suggests that prediction models built using data from other fields cannot be applied reliably to a field that has a limited dataset. Thus, the predictions are largely site-specific.

Table 4-11 Prediction results of RF in the leave-out-one-site analysis (pooled data from five fields).

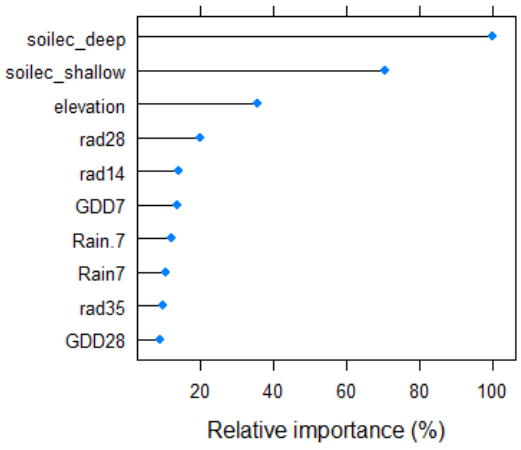
	RF-individual	RF-pooled	Difference	XGBoost-individual	XGBoost-pooled	Difference
				R^2		
Withhold						
NCRS (10 ha)						
2014	0.36	0.35	-0.01	0.16	0.34	0.18
2015	0.35	0.24	-0.11	0.25	0.11	-0.14
2017	0.22	0.19	-0.03	0.21	0.14	-0.07
2018	0.2	0.18	-0.02	0.19	0.11	-0.08
Average	0.28	0.23	-0.05	0.2	0.18	-0.03
Field 2 (5 ha)						
2014	0.11	0.01	-0.1	0.2	0.02	-0.18
2015	0.37	0	-0.37	0.09	0.02	-0.07
2016	0.19	0.01	-0.18	0.01	0.06	0.05
2017	0.1	0.07	-0.03	0.01	0.06	0.05
2018	0	0.01	0.01	0	0.01	0.01
Average	0.15	0.02	-0.13	0.06	0.03	-0.03
Field 3 (14.5 ha)						
2014	0.44	0.19	-0.25	0.15	0.38	0.23
2015	0.69	0.43	-0.26	0.59	0.6	0.01
2016	0.74	0.45	-0.29	0.63	0.68	0.05
2017	0.54	0.5	-0.04	0.51	0.52	0.01
2018	0.09	0.14	0.05	0.09	0.18	0.09
Average	0.5	0.34	-0.16	0.39	0.47	0.08
Field 4 (7.8 ha)						
2005	0.11	0.07	-0.04	0.06	0	-0.06
2007	0	0.01	0.01	0	0	0
2009	0.06	0.1	0.04	0.03	0.15	0.12
2010	0.02	0.04	0.02	0.04	0.02	-0.02
2013	0.27	0.01	-0.26	0.19	0.01	-0.18
2015	0.08	0.04	-0.04	0.05	0.04	-0.01
2017	0.04	0	-0.04	0.06	0	-0.06
Average	0.08	0.04	-0.04	0.06	0.03	-0.03
Field 5 (24 ha)						
2008	0.49	0.57	0.08	0.43	0.38	-0.05
2009	0.54	0.41	-0.13	0.48	0.36	-0.12
2010	0.08	0.03	-0.05	0.1	0.08	-0.02
Average	0.37	0.3	-0.07	0.34	0.27	-0.07

4.4.2 Modelled variable importance using pooled data

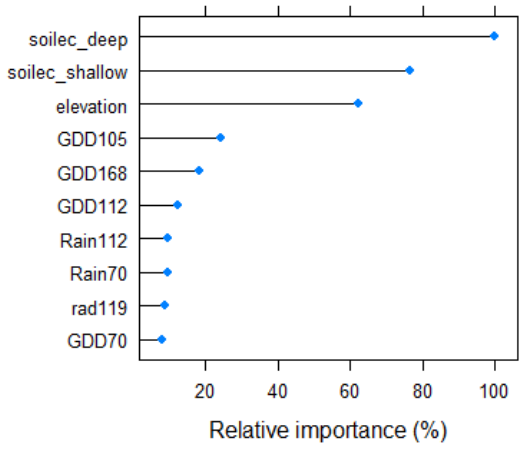
Soil EC deep was consistently ranked as the most important variable in both the RF and XGBoost models using data pooled from five fields in Waikato over years (Figure 4-12). This result is consistent with the hypothesis that areas with higher EC values generally have greater soil water holding capacities and provide plants with more resilience to dry weather conditions.

In the Random Forest (RF) model, the importance of each predictor was calculated by quantifying how much the prediction accuracy (indicated by an increase in the mean square error) will be degraded after permutation using the "out-of-bag" samples (Chapter 3.4.5). Figure 4-12a shows that in the period before V6 (six-leaf vegetative stage – V6), the most important variables in the RF were: "soilec_deep" (100%), followed by "soilec_shallow" (70%) and "elevation" (39%). These were also important predictors in the periods after V6: "soilec_deep" (100%), "soilec_shallow" (79%) and "elevation" (61%) (Figure 4-12b). Compared to spatial effects, the temporal predictors (rainfall, radiation and GDD) had little influence on spatial yield. This result is likely to be constrained by having yield maps from a limited number of years, so is not able to adequately statistically capture temporal yield variability.

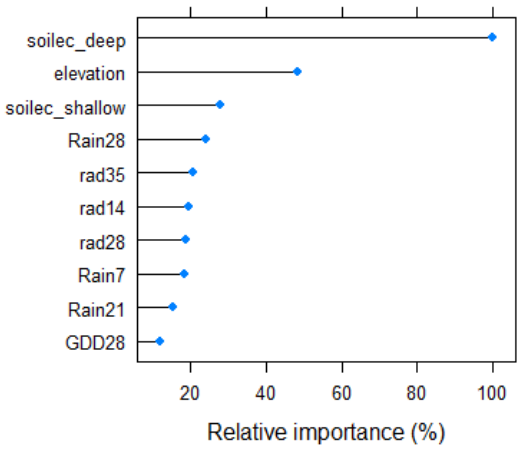
In XGBoost, the variable importance was calculated based on the number of times each variable was split and then divided by the total number of generated trees (Hastie et al., 2009, p. 367). The more a predictor is used to determine key decisions in decision trees, the higher its relative importance. Figure 4-12c, d shows that the most important predictor is "soilec_deep" (100%), followed by "elevation".



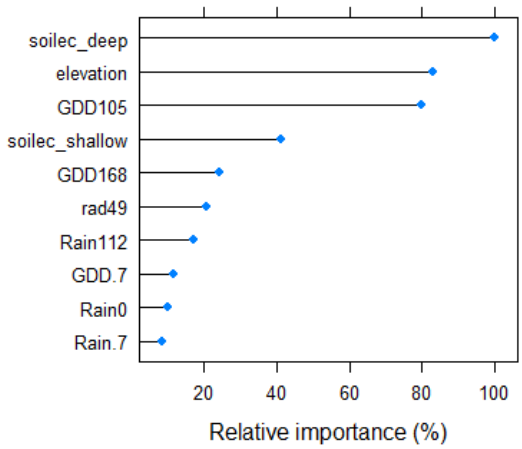
(a) RF (periods before V6)



(b) RF (all periods)



(c) XGBoost (periods before V6)

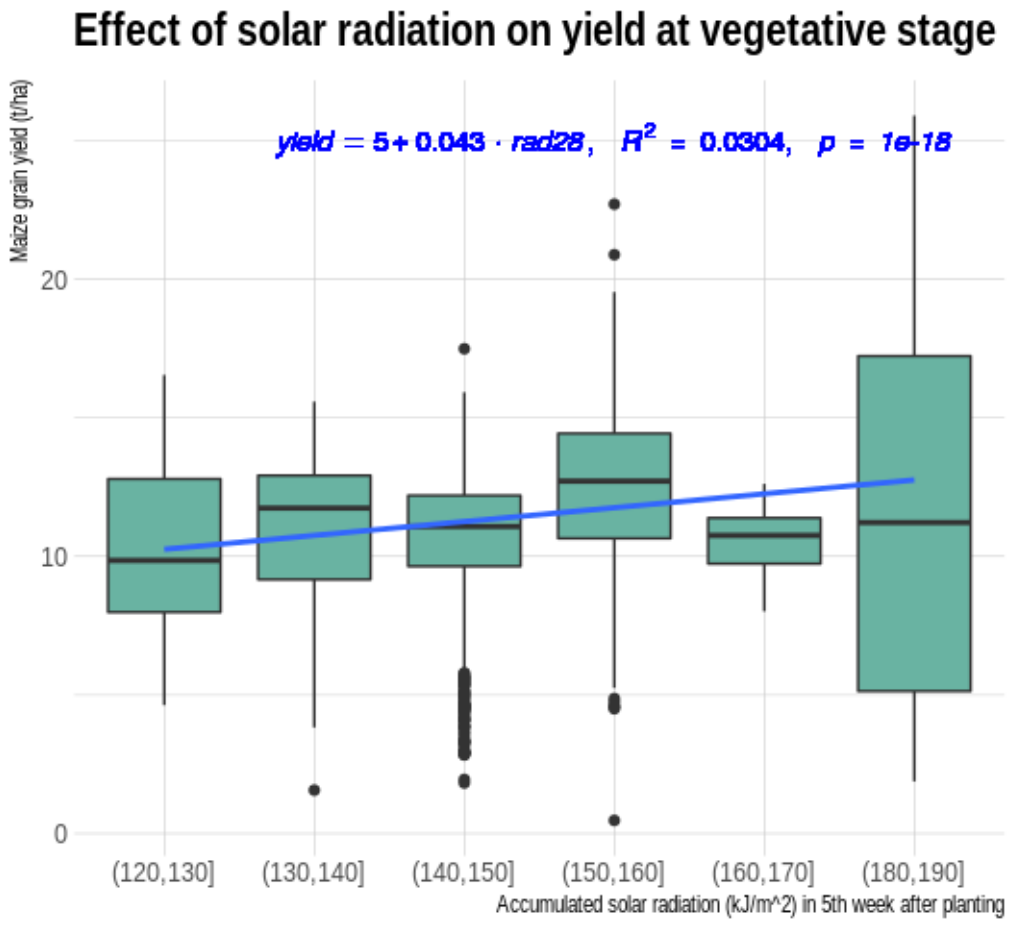


(d) XGBoost (all periods)

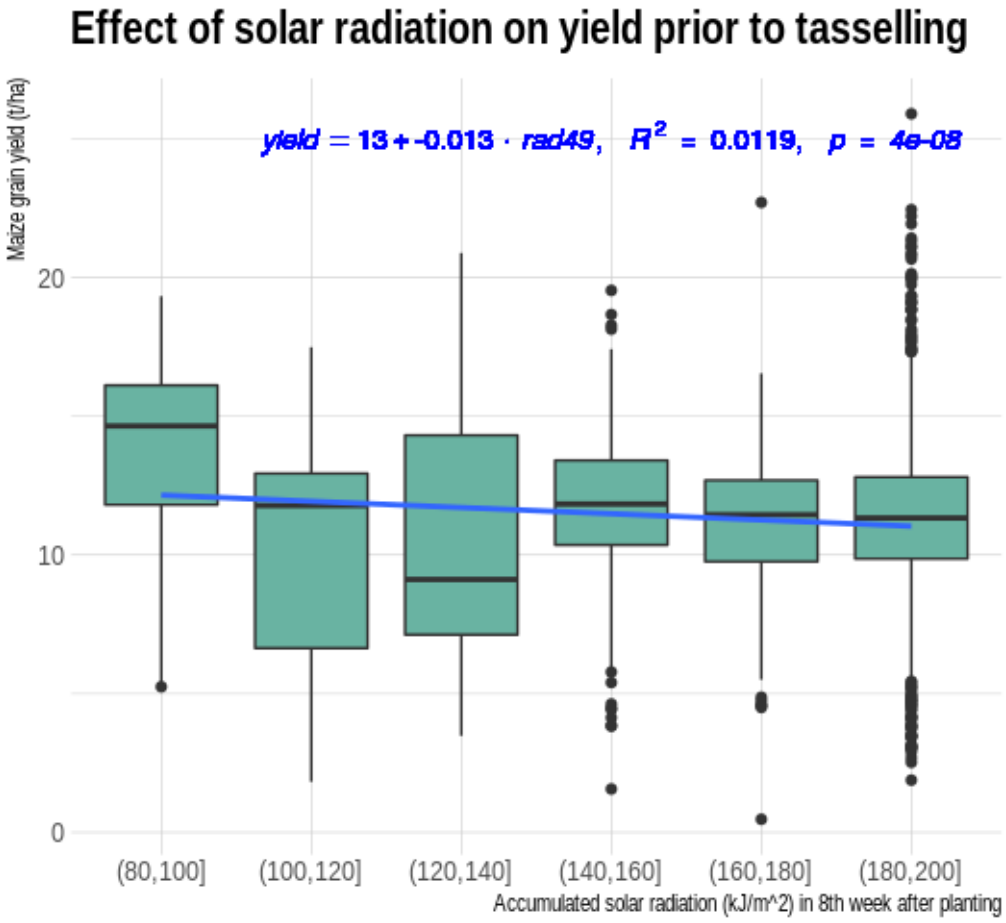
Figure 4-12 Variable importance plot for random forest model (a, b) and XGBoost model (c, d); the importance for each variable was scaled into 0 – 100% based on their relative ranking.

For important temporal yield predictors in the RF model, the accumulated solar radiation in the fifth week after planting was identified as the most important temporal variable (26%). Solar radiation is important for crop growth at various stages of canopy development. The estimated coefficient (Figure 4-13a) suggested that every increase of 1 kJ/m^2 in the accumulated solar radiation during the fifth week may generate an increase of 43 kg/ha in yield (provided there is sufficient moisture), possibly associated with a warmer temperature (the accumulated growing degree days in the fifth week after planting [GDD28])

that is favourable for maize. However, maize-grain yield appears to decline and becomes more variable between sites or between years, with further increases of solar radiation beyond 160 kJ/m^2. A slight decrease of yield was also observed as the increase of solar radiation from 80 to 200 kJ/m^2 approximately two weeks before tasselling, indicated by the negative coefficient of 13 kg/ha (Figure 4-13b). At the early vegetative stage of the crop, the canopy is not fully developed to intercept the radiation. The greater the amount of solar radiation arriving at the plant, the greater the rate of evapotranspiration and hence soil moisture depletion (Ferrante & Mariani, 2018).



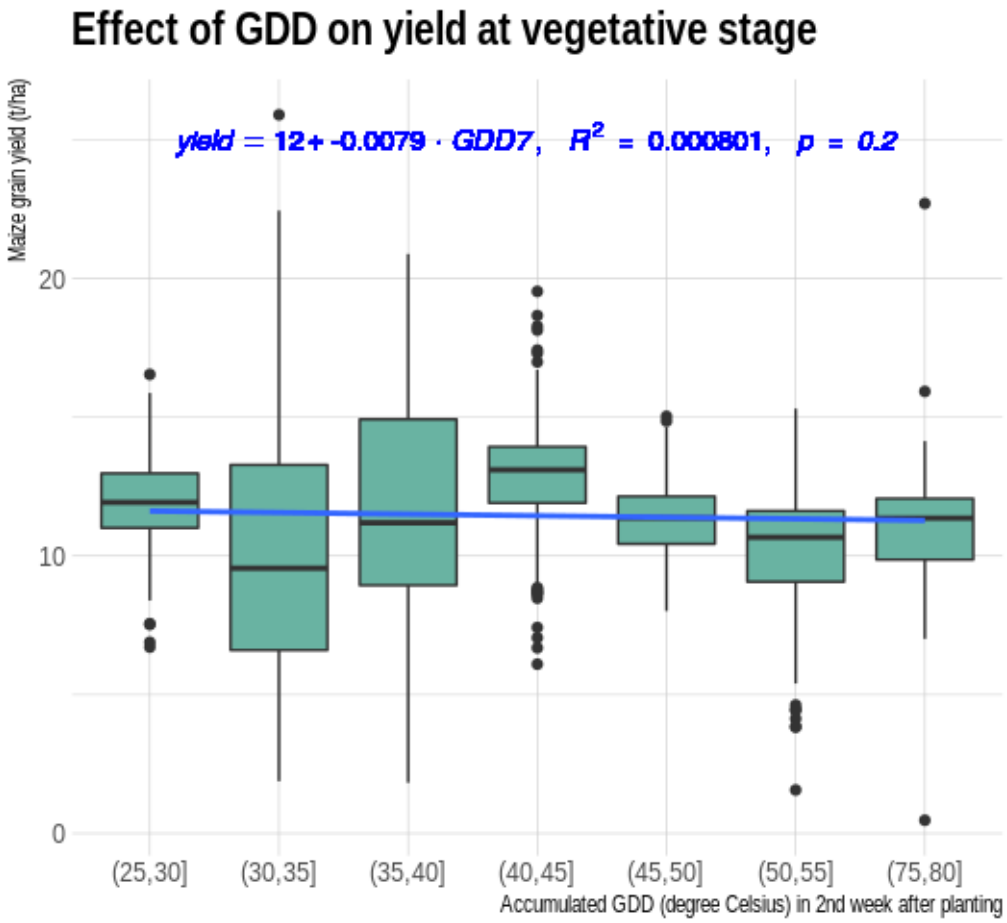
(a) 28 days after planting



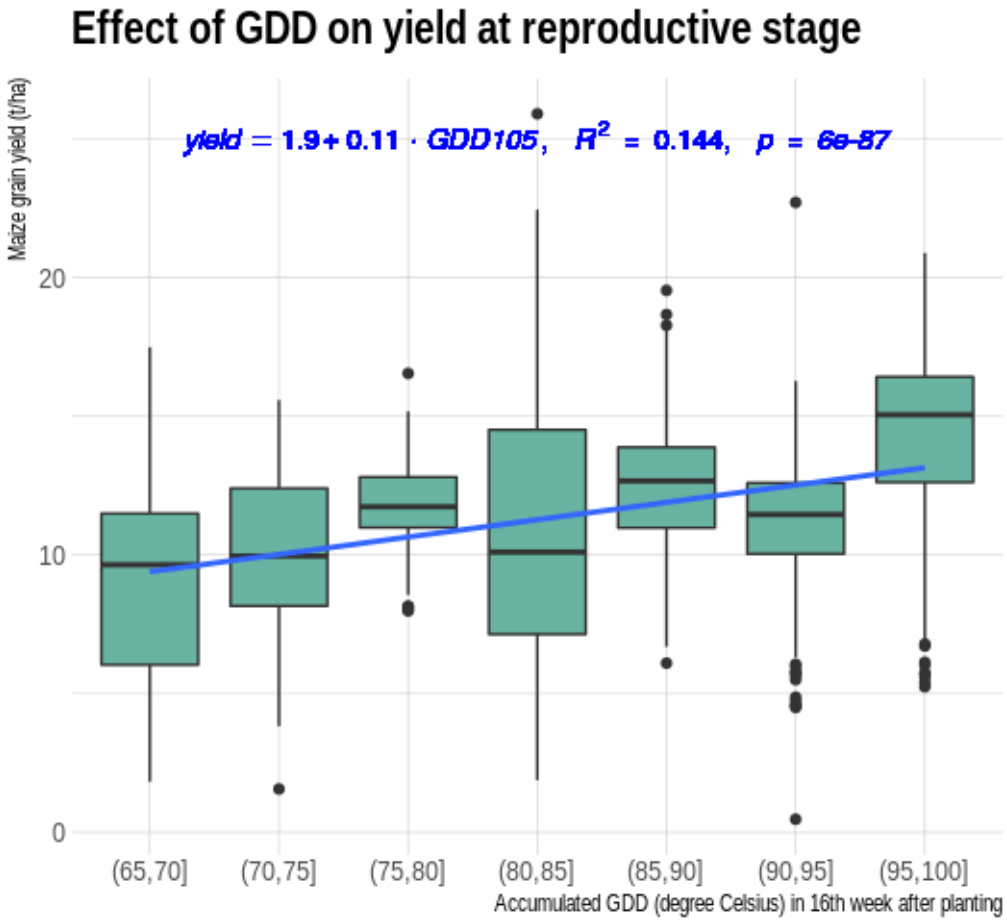
(b) 49 days after planting

Figure 4-13 Yield response to accumulated solar radiation at (a) crop vegetative stage (28 days after planting) and (b) reproductive stage (119 days after planting)

The accumulated GDD during different growth periods also contributed to the temporal yield variability. Figure 4-14a shows that in the second week after planting (V2), accumulated GDD between 40 and 80 degree Celsius tended to produce more stable yield than sub 40 degree Celsius, possibly related to a warmer soil temperature encouraging crop establishment. In the sixteenth week after planting (around early- and mid-February at crop reproductive stage), yield increased with an increase of GDD (Figure 4-14b). The estimated coefficient (0.11) suggests that every increase of 1 degree Celsius in accumulated GDD during this period may cause an increase of 0.11 t/ha in yield.



(a) 7 days after planting

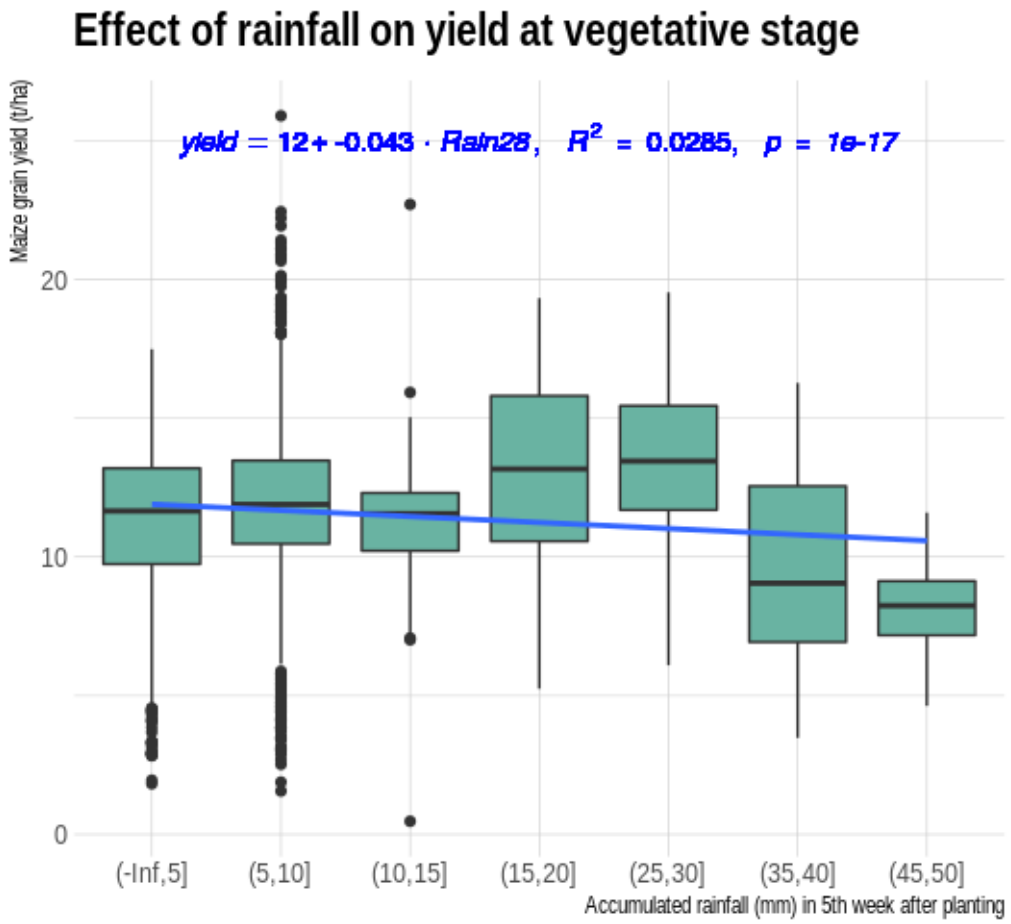


(b) 105 days after planting

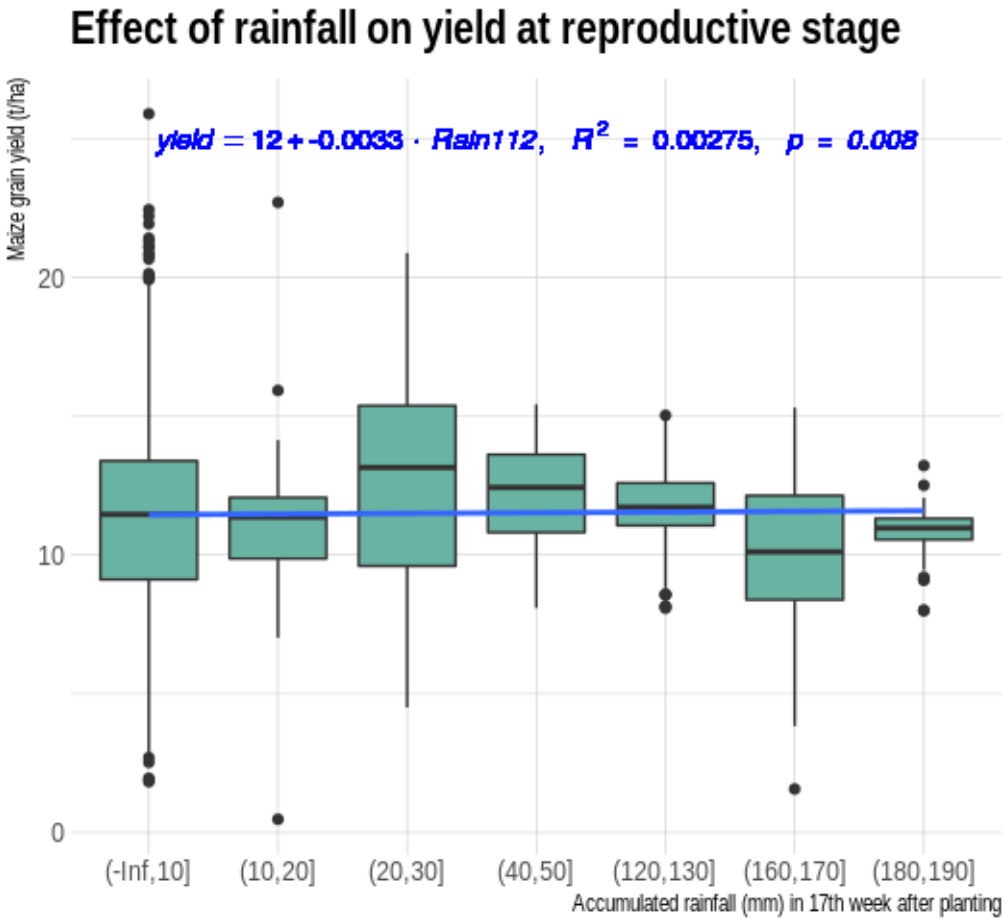
Figure 4-14 Yield response to accumulated GDD at (a) crop vegetative stage (7 days after planting) and (b) crop reproductive stage (105 days after planting)

In the XGBoost model, "Rain28" (accumulated rainfall 28 days after planting, between day 28 - 35) was identified as the most important temporal yield predictor. Figure 4-15a shows that accumulated rainfall in the fifth week (day 28 - 35) after planting (early- to mid-October) of between 15 and 30 mm appears to have produced higher average yields (12-13 t/ha), compared to 7-8 t/ha yields produced when the rainfall was less than 15 mm during week 5. Rainfall in early November before applying midseason N fertiliser was important for providing sufficient moisture for the crop. However, as the 28-35 days rainfall increased above 35 mm, the yield dropped dramatically, possibly related to starter N leaching below the shallow root zone (30 cm (Abendroth et al., 2011)). The negative coefficient of 0.043 ($p<0.05$) suggests that every

increase of 1 mm in the accumulated rainfall at early vegetative stage may decrease yield by 43 kg/ha. There has been an increase in yield variations with an increase of rainfall at this stage, suggesting that yield responded differently between different fields and soil types. Figure 4-15b shows that the accumulated rainfall 112 days after planting at crop reproductive stage also contributed to yield, despite having less significance in both models. These predictor importance analyses help identify important factors to further fine-tune yield prediction models.



(a) 28 days after planting



(b) 112 days after planting

Figure 4-15 Yield response to accumulated rainfall at (a) crop vegetative stage (28 days after planting) and (b) crop reproductive stage (112 days after planting)

4.5 Summary

This chapter presented the following results:

- Section 4.2 addressed Objective 1 (To develop a filtering algorithm to improve maize yield mapping precision.)

Spatial filtering applied to the yield monitor data over the four years (2014, 2015, 2017 and 2018) removed between 6 to 40% points from the original yield datasets (Table 4-1). The yield data for most years (2015, 2017 and 2018) had less than 6 to 12% removed. The highest number of points removed (40%) was associated with the yield in 2014. The spatial filtering method that was applied reduced the RMSEs of kriging for all available years (2014, 2015, 2017 and 2018). These results suggest that the method developed in R programme was effective for improving the accuracy of the yield maps. The filtered yield data were used throughout the subsequent analyses.

- Section 4.3 addressed Objective 2 (To identify appropriate spatiotemporal yield predictors by examining historical yield maps, delineating subfield management zones and undertaking zonal soil sampling.)

The historical yield data and soil EC collected from NCRS were mapped and delineated MZs. A good statistical agreement (Table 4-5) was found between yield productivity zones and crop reflectance zones, which suggests that the spectral images derived from Sentinel-2 can be a proxy of spatial yield for delineating management zones when yield maps are absent. Soil sample cores were taken from each yield productivity zones and measured for soil texture: sand, silt, and clay. The soil EC data were positively correlated (R^2=0.69; profile average) to the clay fractions and the silt fractions (R^2=0.47; profile average) at the soil surface. The nonlinear quadratic models were able to explain the relationships between the spatial yield of each year and soil EC (R^2=0.44-0.92). These results reaffirm that soil EC can be used as a proxy for soil texture to explain the cause of the spatial yield variability.

- Section 4.4 addressed Objective 3 (To determine the viability of predicting dynamic maize yield at the subfield spatial scale using supervised machine learning algorithms.)

The statistical model hyperparameters were optimised using the grid search and ten-fold cross-validation. The nonlinear models (FFNN, CART, RF, XGBoost and Cubist) produced more accurate yield prediction (R^2 = 0.36 - 0.72) than the SMLR model (R^2= 0.20 - 0.72), see Table 4-9. RF, XGBoost and cubist (R^2= 0.44 - 0.72) produced relatively better prediction results than the SMLR and FFNN. The accuracy of CART was close to FFNN (R^2= 0.52 - 0.60). However, the yield prediction for the individual year (Table 4-10) was generally poor with a small R^2 produced. Better predictions were generally associated with greater yield variability within-field in some years indicated by the coefficients of variation (CV%). RF, XGBoost and Cubist produced better results than the other trained models. RF produced the highest model fits (average R^2 = 0.08 - 0.50), followed by XGBoost (average R^2 = 0.06 - 0.39) and Cubist (average R^2 = 0.03 - 0.19). Then these modelling analyses were repeated using data pooled from different fields to determine if the models built using data collected from other sites can be applied to predict yield for a specific site that has no or limited data (Table 4-11).

To provide insights into managing spatiotemporal yield variability, variable importance analysis for RF and XGBoost was conducted to identify important yield predictors. In addition to spatial variables such as soil EC and elevation, several variables such as accumulated solar radiation (in 5th and 8th week after planting), GDD (in 2nd and 16th week after planting) and rainfall (in 5th and 17th week after planting) were proven to be useful contributors to the prediction of yield.

Chapter 5 Discussion

5.1 Introduction

As yield is the driver of revenues for arable farmers, establishing zones where revenues and expenses can be managed or generating site-specific crop management zones (MZs) is a key concept in precision farming. In New Zealand, yield monitor data have frequently been collected on-farm from harvests over time, but the monitors have been rarely calibrated. The commercial uptake of precision farming practices, such as calibrating yield monitors has been limited due to a lack of proven commercial benefits. Therefore, little has been done to find methods to delineate zones from a variety of spatiotemporal information for crop management.

The research outlined in this study aims to determine if the process of delineating site-specific MZs in maize crops can be improved by modelling spatiotemporal interactions between spatial and other complementary factors affecting yield. This discussion chapter discusses the findings of the study within the context of addressing the objectives posed in Chapter 1:

- Section 5.2.1 addresses Objective 1 (To develop a filtering algorithm to improve maize yield mapping precision).
- Section 5.2.2 addresses Objective 2 (To identify appropriate spatiotemporal yield predictors by examining historical yield maps, delineating subfield MZs, and undertaking soil sampling).
- Section 5.2.3 addresses Objective 3 (To determine the viability of predicting dynamic maize yield at the subfield spatial scale using supervised machine learning algorithms).

The discussion is broken into two main sections: 1: Provides more technical evaluation of the methods and analysis undertaken and reviews the findings observed in a broader scientific context and 2: A more general discussion that describes how these findings can benefit the arable industry.

5.2 Technical discussion and evaluation of methodologies

5.2.1 Effectiveness of the spatial filtering algorithm

The results from the filtering analyses (section 4.2) showed that the customised spatial filtering programme removed 6%-40% of the observations (586 to 4581 points) for the study field in each year. These numbers were like that from the previous study by Sudduth and Drummond (2007), which reported 13% to 27% (2510 to 8325 points) removal from the maize and soybean yield data collected from six fields, using the Yield Editor they developed. Spekken et al. (2013) removed 29% of the raw data points using their proposed spatial filtering method and compared this with the method based on traditional statistical upper and lower yield limits. However, further evaluation is still required to determine how these filtering programmes could influence the accuracy of yield maps, Maldaner et al. (2018) combined previous methods reported (global, anisotropic and isotropic filtering) and applied this approach to three Brazilian soybean fields. However, it is difficult to develop anisotropic filtering without a correctly recorded GPS timestamp, which may be missing in the data collected in New Zealand as reported in this thesis. In Maldaner et al. (2018)'s study, relatively large percentages of the data (around 20 to 40%) were removed. The nugget/sill ratios were reduced from 0.86-0.95 to 0.54-0.58 with the range of spatial dependence reduced from 196-5,000 m to 60-339 m. However, these results were different to that reported in Figure 4-1 (0.18 – 0.91 to 0; 106-390 m to 13 to 59 m) because of different crops and environments.

In addition to the geostatistical analyses such as investigating the nugget/sill ratios, this thesis determined how spatial filtering affected the accuracy of spatial prediction (interpolation) of yield data using ordinary kriging with several common variogram models (spherical, exponential, gaussian and Matérn). The results (Table 4-2) reaffirmed that the filtering algorithm was effective at improving the kriging prediction for mapping spatial yield, as indicated by the large reduction of prediction errors for some years (for example, RMSE was reduced from 4.41-4.59 to 0.96-1.43 for the interpolated yield in 2014, and 6.61-7.27 to 1.34-1.43 for the interpolated yield in 2015, see Table 4-2). These results were an improvement compared to the unfiltered data and the errors were far more acceptable.

Very little ground validation was conducted in these analyses on the effect of spatial filtering, common errors in geospatial yield monitor data were reviewed in section 2.3.1.1.1, concentrating on the causes of errors, and how to mitigate these. Previous work shows that the errors are commonly caused by systemic limitations (such as flow delay) and human mistakes such as misuse of cut width setting. This 2014 data set had more points removed than the other years suggesting that it may contain more unusual values due to inadequate sensor calibration (Luck & Fulton, 2014). The unrealistically high yields such as 98.0 t/ha observed in 2014 and 101.5 t/ha in 2015 may have been caused by abrupt vehicle stoppage. The ground travel velocity was reduced to near 0 km/h in a short period while the system was still conveying the grain across the sensing plate and therefore the unrealistically large yield values were generated. The yield values of zero found in the data were likely caused by the process of emptying the conveyor belt, when the harvester turns around at the end of a row with the header down but not harvesting any crop (Colvin & Arslan, 2000). In this thesis, filtering was undertaken to mitigate the impact of erroneous data values, which may cause biases in subsequent multivariate modelling analysis (section 4.4).

5.2.2 Effectiveness of spatial yield predictors

5.2.2.1 Examining spatial and temporal yield variability

The historical yield maps (2014, 2015, 2017 and 2018) generated from field data (section 4.3.1) have been analysed. It can be seen from the 2014 yield map that there was a strong spatial "banding" pattern across the centre of the field which followed that of the stream on the northern boundary. The spatial "bands" that produced a lower yield ranged from 4 to 7 t/ha, compared to the other parts of the field, which mostly produced a higher yield ranging from 10 to 16 t/ha. The occurrence of this "banding" pattern coincided with an extended dry period from late-February to mid-March at a critical reproductive stage in plant development (Figure 4-3), which may have caused plant moisture stress in some locations within the field. Since the field received no irrigation and had been uniformly managed with the same amount of seed and fertiliser, these results suggest that inherent soil moisture variation may have caused the

spatial yield variability, in conjunction with the lower than average seasonal rainfall. In a drier season, the crop growing in the coarser-textured soils would be more susceptible to moisture stress, compared to the crop growing in finer-textured soils which are more likely to have higher water holding capacities (Hewitt, 2004). In comparison to my research, Blackmore (2000) investigated spatial and temporal yield variability using yield data collected over six years from a small 6.7 ha field with winter wheat and oilseed rape rotation in eastern England. Rather than soil texture variations, he found the relatively low- and unstable-yielding areas were mostly located on the field edges (headlands) and generated lower gross margins compared to the centre of the field. A similar effect of field headlands on spatial yield has also been found in New Zealand based on previous yield mapping analyses on other maize fields (Holmes & Jiang, 2018).

The coefficients of variations (CVs) for the multiple-year yield maps were calculated (Figure 4-4) to determine the stability of that spatial yield pattern over the four years. The results show that the higher the CV value, the more temporally unstable the yield was over the years. 80% of the areas had a CV between 10% and 40%. The CV threshold that divides the field into two equal sizes was at around 18.5%. This unstable yielding rate was associated with low-lying areas. This could indicate more susceptible to slower drainage once the perched water reaches the root zone in rainfall events. The area around the study site has been extensively studied (Lowe, 2010) and consists of the Horotiu (Allophanic, Typic Udivitrand) and Te Kowhai (Gley, Typic Ochraqualf) soil complex. The nearby stream has degraded the landscape features adjacent to its flow, leaving remnant lenses of the poorly drained Te Kowhai silt loam. The persistence of soil moisture conditions close to field capacity caused by poor drainage is likely to deplete the soil oxygen content critical for maize root respiration and affect the plants' ability to absorb nutrients (Saglio et al., 1983). Similarly to my results, Blackmore et al. (2003) investigated temporal yield variance over six years for four fields in England that had grown malting barley, winter wheat and oilseed rape. In one of the fields, they found a relatively high and unstable yielding MZ in a low-lying deep soil area next to an open drain, which may have caused large temporal yield variations from one year to another, depending on average seasonal variations in rainfall.

Lark and Stafford (1997) used a different statistical approach from that used in this thesis. They used fuzzy c-means clustering to group yield into several clusters and then interpreting local yield variation using the fuzzy c-means membership values. The membership value is the association between one sampled yield data point and each clustered yield zone, to help evaluate the statistical confidence of clustering of the yield data. Their analysis conducted on a 6-ha field that grows winter barley during 1993 to 1995 in the UK, suggested that unstable yield may be attributed to an observed poor drainage status recorded over the past season within the field. The results in this thesis support the effect of field elevation on spatial yield in relation to temporal seasonal rainfall and suggest that the use of elevation derived from an RTK GPS on a tractor is a useful predictor for yield.

5.2.2.2 Examining soil EC and texture variability on spatial yield

5.2.2.2.1 Soil EC mapping and sampling resolution

To examine the effect of soil EC on spatial yield, the soil EC data were interpolated using ordinary kriging with adequate fitted variograms. Table 4-4 presented the properties of these variograms and showed that the range of spatial dependence on the distances that separate two locations were similar for the soil EC measurements (38 m and 41 m) for shallow (0-30 cm) and deep mode (30-90 cm) respectively. Similar findings were derived from a previous study in New Zealand by Hedley and Yule (2009). They investigated spatial soil texture variations for a 35-ha irrigated maize field in Palmerston North, intending to predict soil moisture distribution. They interpolated the soil EC data with 14,152 points measured from a Geonics EM38© sensor and delineated this data into zones for further soil sampling. Their results showed a similar range of spatial dependence (33 m) to this study (38 m and 41 m).

The range of spatial dependence derived from soil EC maps is useful for determining the optimal soil sampling intervals. As a rule of thumb, a sampling interval of less than half-of-the-range of spatial dependency from the ancillary data should be considered to compute a variogram reliably for kriging (Kerry & Oliver, 2003). If this rule is to be followed, to develop a variable rate fertiliser application strategy,

soil samples should be taken from less than 19 m apart for mapping soil fertility via kriging, for prescribing a map-based variable rate application of starter fertiliser, a sampling resolution (cell size) of 19 m is applicable, because it is larger than the full width (6 m) of a typical maize planter in New Zealand. However, for a mid-season fertiliser application, the full bout width of a fertiliser spreader in New Zealand is approximately 20 m, which is about the same as the 19 m optimal interval for soil sampling. Producing soil fertility maps of less than 19 m resolution may incur unnecessary sampling costs if the technology is not capable of applying inputs at the resolution that is required. For evaluating the potential effect of VRS strategies on spatial yield, spatial maps with 6 m resolution were produced as predictors for yield.

5.2.2.2.2 Effect of soil texture on EC and water holding capacity

In this study, soil EC was collected from the field using a direct EC sensor (Veris MSP-3). Both soil EC maps scanned at two depths showed a clear spatial "banding" pattern with local EC values ranging from 0 to 3 mS/m, which indicates potential coarser-textured soils. The researcher thus speculated that this pattern originated from geological processes such as the deposition of fine alluvial materials from nearby streams during flooding events. To correlate EC with soil texture, soil samples were collected at six locations (based on yield productivity; described in section 3.3.5) across the field using a soil corer at depths of 5-10 cm, 10-15 cm, 15-20 cm, 20-25 cm and 25-30 cm. Soil samples collected from the field were analysed for texture (sand > 63 microns, silt 2 - 63 microns, and clay < 2 microns) and replicated. The results of the analysis of these soil samples show that soil EC was positively correlated with the fine particles associated with the average silt and clay fractions in the topsoil of 5 – 30 cm (R^2 = 0.47 and R^2 = 0.69, respectively). These results are consistent with a previous study by Sudduth et al. (2005) in the north-central US, which found the highest correlation (R^2 = 0.37 - 0.63) between EC and topsoil clay content across 12 fields. In New Zealand, Hedley and Yule (2009) reported that the zones classified by higher EC values were associated with higher silt (30.8 to 61.3%) and clay content (8.3 to 26.8%) for the top 60 cm soil, which is similar to that in Table 4-7 (31% to 69% for silt; 12% to 27% for clay). They then determined the available

water holding capacity (100 to 190 mm/m) and irrigation trigger for each soil zone and develop real-time variable rate irrigation strategies. Similar relationships between soil EC and soil texture between zones were also derived from Hedley et al. (2010)'s study on a 22-ha maize field in the sand country region of Manawatu. The results reported in Table 4-7 suggest that large spatial soil texture variations not only exist on a 20-40 ha field under centre pivot irrigation but also exist in an arable field as small as 10 ha, which highlights the applicability of variable rate application on small fields.

A guide for irrigation provided by FAR (2010) recommended that the plant available water should be at least 10 times greater than the daily evapotranspiration to maintain a steady uptake of water for the optimum yield. Based on the soil particle size distribution measured at six sampled locations (section 4.3.4.1), it is estimated that the maximum amount of the plant-available water at these locations ranges from 136 mm/m to 180 mm/m, which is similar to Hedley and Yule (2009)'s previous estimation (100 to 190 mm/m) for their research field, even though the soils types are different between the studies. As a rule of thumb, the crop can extract a maximum of 10% of the available water per day from the root zone (FAR, 2010). Theoretically, given potential evapotranspiration of 5 mm per day in summer, it would take 13 days for maize yield reduction to occur in the soil at the field capacity with 180 mm/m of plant-available water with no rainfall recorded. These results suggest that crop moisture stress during a period after the V6 is more likely to occur when the nil rain condition persists for approximately two weeks, because of an inherently larger plant available water capacity.

5.2.2.2.3 Effect of topsoil texture on spatial yield

In the study site reported in this thesis, maize seeds were planted at 5 cm depth, with polymer-coated urea added 2.5 cm below seed placement. According to Genetic Technologies Limited (2016), 20% of the total seasonal water required by the crop occurs in the first five weeks of crop establishment. From planting through to the V6 stage, the rooting depth of a maize crop is estimated to reach 30 cm deep. During periods of low rainfall, coarser textured horizons (in the top 30 cm soil), hosting the bulk of the

root mass, may have limited ability to meet the water demands for the growth onward (Hedley et al., 2010; Stadler et al., 2015), particularly in a non-irrigated system. Severe and prolonged water stress in maize plants during the early vegetative stages may cause irreversible damage to the structure of the photosynthetic membrane, resulting in lower chlorophyll content and possibly a 25%-30% reduction in yield (Cakir, 2004; Denmead & Shaw, 1960), even if sufficient water supply is present during the remaining growth period (Song et al., 2019). Additional urea fertiliser is typically broadcast at the V4 to V6 stage (5-6 weeks after planting; mid-November) to support the growth (Figure 5-1). However, the applied nutrients are not be solubilised for uptake if there is little soil water stored at the 0-30 cm depth. Compared to the V4-V6 stage, at a later growth stage (tasselling; mid-December; 75-85 days after planting), the root system is more deeply established in the soil (>70 cm) and can access a larger volume of plant-available water, unless dry seasonal conditions persist for more than 13 days.

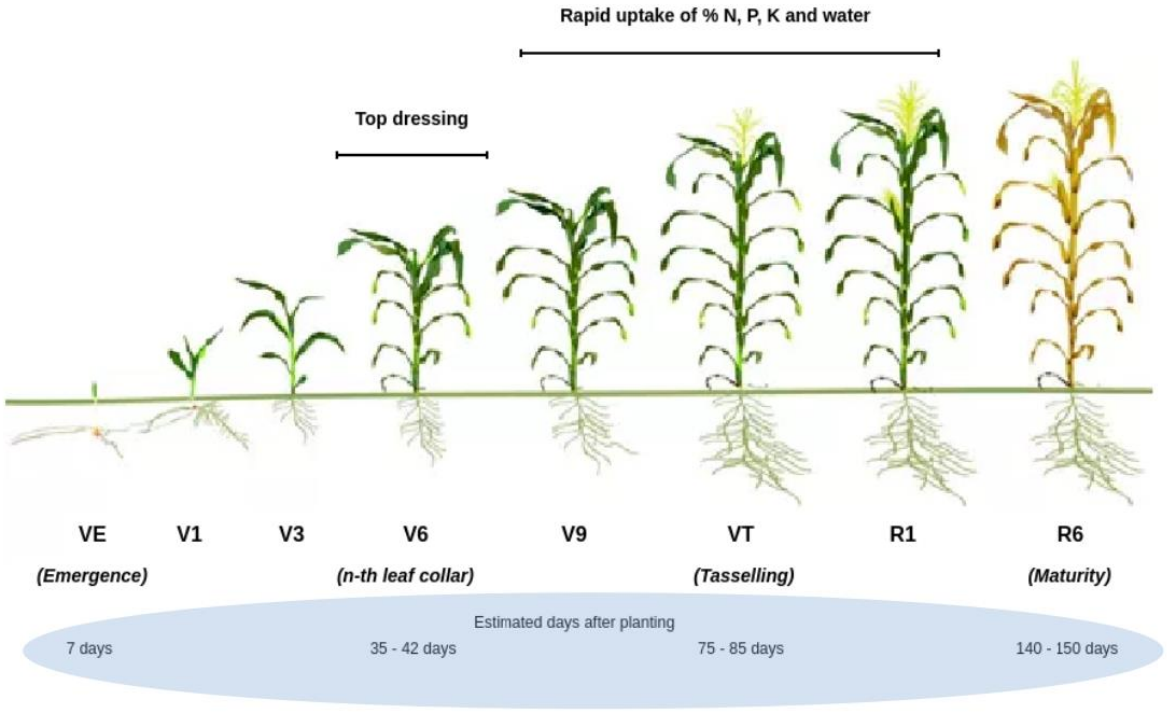


Figure 5-1 Maize growth stages in New Zealand and the estimated days after planting required (the rates of growth depends on the environmental conditions). Image modified from Genetic Technologies Limited©

Literature shows that yield often increases with an increase of soil EC because this data suggests better water holding capacity and lower risk of moisture stress during crop growth (Hedley et al., 2010; Stadler et al., 2015). However, yield may then decrease after optimal soil EC due to impedance of drainage and the onset of anaerobic soil conditions (Corwin & Lesch, 2003).

In this study, quadratic models were fitted for each year to investigate the relationship between EC and historical yield maps. Table 4-8 showed that the model had a good fit for 2015 and 2017 (R^2 = 0.76 and R^2 = 0.92, respectively) but a poorer fit for 2014 and 2018 (R^2 = 0.44 and R^2 = 0.56, respectively), suggesting a temporally variable yield response. Management factors (such as the timing of mid-season fertiliser application), as well as variable rainfall patterns, could play an important role in determining yield potential. This variability was consistent with a previous study in the north-central US by Kitchen et al. (2003). Their study modelled the relationship of soil EC, topographic features (elevation, slope and curvature) and crop yield for three contrasting soil-crop systems. Their quadratic models explained 7%-

51% (an average of 21%) of the maize-grain yield variation using the EC data alone. Some improvements for the models were produced (an average of 32%) when combining soil EC and elevation for the non-irrigated fields in Kansas and Missouri but only slight improvement (by around 1-2%) for the irrigated field in Colorado. In Table 4-5, the combination of EC and elevation maps reported no significant improvement on correlating to the yield zones, because there is little variability of elevation within-field at NCRS (a gradual decrease of elevation by approximately 4 m for over 400 meters).

Kitchen et al. (2005) delineated productivity zones using a combination of soil and elevation maps and compared these with yield productivity zones delineated using historical yield maps over seven years for two claypan soil fields in US midwest planted in a maize-soybean rotation. The historical yield maps (between 1993 and 2002) were grouped into ``deficit", ``optimal" or ``excessive" years based on the spatial yield in specific landscape positions (foot-slope and side-slope) and in relation to the quartile threshold of monthly rainfall over 58 years. The grouped yield maps were then delineated into the respective yield zones for these temporal conditions. Kitchen et al. (2005)'s study reported a 51–68% agreement (with kappa coefficient κ = 0.21 - 0.43) between multiple-year yield zones and soil MZs. The highest agreements (58-60%; κ = 0.29 – 0.34) were found with the yield zones derived from the "deficit" years (the yield in the eroded side-slope was less or equal than 95% of the field average), whereas relatively lower agreements (41-46%; κ = 0.07 – 0.11) with the yield zones derived from the "optimal" years (the yield in both eroded side-slope and foot slope between 0.95 and 1.05 of the normalised field average yield). The results reported in this study (53-63% agreement and κ of 0.04-0.27, Table 4-5) was consistent with Kitchen et al. (2005)'s findings. The low κ indicates low statistical confidence of the agreement between yield and soil MZs, which is likely to be associated with the high number of pixels of the zone maps (N = 4,346; Table 4-4).

5.2.2.2.4 Reaffirming effect of soil texture on yield using satellite images

The purpose of introducing the satellite images into section 4.3.2.3 was to determine if these freely accessible, relatively coarser-resolution images (10 m) can reflect the influence of soil texture variations

on maize yield in this study. There was a high agreement of 71% (κ = 0.42) between yield productivity zones and crop reflectance zones (Table 4-5), which suggests that these multispectral images (Sentinel-2 images captured in February at the crop reproductive stage) could be used as proxies for spatial yield for delineating MZs, especially when yield maps are absent. Although different satellite images were employed, the result in this thesis was consistent with Georgi et al. (2018)'s study, which developed an automatic algorithm for delineating crop reflectance zones using fine spatial resolution RapidEye satellite multispectral images (5 m). The crop reflectance zones derived from the NIR bands of all automatically selected images were able to delineate significant yield differences (p<0.05) related to soil texture variations within-field for different crops (wheat and canola) on different fields. Table 4-6 shows that medium-resolution crop reflectance zones are also potentially effective at differentiating the yields of maize crops, which could support the further uptake of soil EC mapping on small arable fields for precision farming in New Zealand.

5.2.3 Predicting spatial yields at the subfield level

5.2.3.1 Modelled performance

To assess the performance of the models (SMLR, FFNN, CART, RF, XGBoost and Cubist) that have been trained (section 4.4.1) and to compare different algorithms, the trained models were assessed using a subset of the data withheld from the original dataset. Around 20 - 72% of the yield variations were explained by the models. SMLR was not very useful for predicting yield (average R^2= 0.20 - 0.52), consistent with a previous study by Sudduth et al. (1996). Slightly better accuracies in this study (average R^2=0.42 - 0.72) were produced by FFNN and CART than the linear regression model. These accuracies were reasonable for predicting spatial yield, as R^2 between 0.21 to 0.74 were reported previously (Drummond et al., 2003). For CART, less accurate results were expected because the model predicts discrete values (i.e. the predicted value is the average value for each data subset), instead of continuous values as in more complex ensemble tree methods (Miller et al., 2016). For FFNN, less accurate results were also expected

because the optimised FFNN model was functionally similar to multiple regressions when only a few hidden neurons were required. Further improvement on the prediction accuracy of spatial yield by the neural network may be possible with more complex architecture (deep neural network with two or more hidden layers) (Orimoloye et al., 2020). However, training these neural networks can be computationally time expensive. The time used for training the model could vary from hours to weeks depending on the structure of the neural network (e.g. the number of hidden layers, neurons) and the optimisation methods, which increases the difficulty of processing spatially dense precision farming data. Better accuracies (R^2 = 0.45-0.72) were produced by the ensemble tree models (RF, XGBoost and Cubist), possibly because of continuous outputs rather than the discrete outputs produced by CART.

To assess how well a model predicts unseen data (i.e. yield from an additional harvest), the leave-out-one-year analysis was undertaken by withholding one year of yield, as the test set. However, the prediction accuracies of all the models for individual years were generally poor given the relatively limited number of years (2014, 2015, 2017 and 2018), which were unable to capture the range of temporal effects such as rainfall on yield. The XGBoost and RF models tended to produce a better prediction of yield. It is suggested that XGBoost and RF both have potential in handling data with multiple spatiotemporal interactions. These should be promoted to predict yield at a subfield level, although at greater computational costs than other models. Compared to the RF model, in which there is no pruning for the trees generated, XGBoost allows its trees to be pruned and thus tends to avoid overfitting. However, compared to RF, which employs “bootstrapping” (random sampling with replacement) to generate new datasets to avoid biases from individual observations, XGBoost is more susceptible to “noise” (i.e. unrealistic values in any of the input datasets). Given the potential inconsistency in the yield sensor calibration methods, it would be difficult for XGBoost to capture that multivariate relationship if the historical data of some years do not represent the actual yield level (see Chapter 2 section 2.3.1.1.1).

These results were consistent with the study of Drummond et al. (2003). The authors predicted maize and soybean yield on three sites in the US state of Missouri (ranging from 13 to 36 ha in size) using an FFNN

model with several soil fertility parameters (soil pH, OM, phosphorus, calcium, magnesium, potassium) and topographic inputs (elevation, slope). The FFNN model generally provided better statistical predictions (an average of 45% of the variation, ranging from 21 to 74%) in a multiple-year analysis than the other two models (SMLR, which explained an average of 31% of the yield variation, range 14% to 65% and Projection Pursuit Regression). However, in the leave-out-one-site-year analysis, significant prediction errors were produced for the individual site year due to severe overfitting. Consequently, most climatological variables were excluded except for the total rainfall during the plant reproductive phase (days 51–110 after planting). It is concluded that a much larger set of climatologically unique sites and years, would be required for these models to be used in a predictive manner. However, the minimum number of sites and years data required to produce a reasonable prediction of spatial yield is unclear. The results may be distorted with additional years of yield data, in the leave-out-one-site-year analysis, if the consistency of historical management actions and yield data quality is not known.

A similar statistical approach was applied in a study by J. Liu et al. (2001) in Urbana-Champaign, Illinois, US. This used an FFNN model for predicting maize yield based on a range of climate (Growing Degree Days (GDD) and monthly rainfall) soil factors (soil pH, phosphorus, potassium, organic matter), management factors (nitrogen fertiliser) as inputs in their FFNN model. By doing this they attempted to predict yield on small plots with different fertiliser treatments in the Morrow Plots of the University of Illinois. In comparison, this study attempted to predict spatial yield from commercial cropping fields. However, only one model (FFNN) was tested in the study by J. Liu et al. (2001), whereas in this study, the ensemble tree models (RF, XGBoost and Cubist) reported better predictions of spatial yield (average $R^2 = 0.44 – 0.72$) for data with multiple spatiotemporal interactions than from FFNN (average $R^2 = 0.39 – 0.72$). These tree-based models that are trained with field data could provide a statistical basis to help delineate dynamic yield zones in New Zealand, compared to the static zones evaluated previously in US midwest (Kitchen et al., 2005), for optimising management inputs within-field.

5.2.3.2 Potential improvements to the yield prediction models

5.2.3.2.1 Inclusion of soil moisture data

Maize crop yield was generally influenced by seasonal rainfall at several critical stages. The analysis of relative variable importance (Chapter 4.4.2) suggested that the accumulated rainfall in the fifth week and the seventeenth week after planting were identified as influential factors for the yield in the XGBoost model (Figure 4-12c, d). Figure 4-15a shows that accumulated rainfall in the fifth week after planting (between day 28 - 35; early- or mid-October) of between 15 and 30 mm appears to have produced higher average yields (12-13 t/ha), compared to 7-8 t/ha yields produced when the rainfall was less than 15 mm, based on data pooled from five fields over the years. The accumulated rainfall in the fifth week after planting provides enough moisture for dissolving N fertiliser to nitrate that can be readily taken up by plants. Adding a small amount of starter fertiliser at planting is traditionally recommended for growing maize in a temperate climate regime as it benefits the growth of the primary nodal roots for water and nutrient uptake and lowers the risk of uneven stands of maize (Mascagni et al., 2007). It is also a more effective way of placing other valuable, and less leachable nutrients such as P and K in the top 10 cm soil, than broadcasting via a fertiliser spreader.

An excessive amount of rain at the seedling emergence stage may cause nitrate leaching from the shallow root zone (Butts-Wilmsmeyer et al., 2019), especially with coarser-textured soils, which have a lower water holding capacity and (if free-draining) have a higher throughput of pore-water volume. Figure 4-15a shows a negative estimated coefficient of 0.043 ($p<0.05$), which suggests that every increase of 1 mm in the accumulated rainfall at early vegetative stage may decrease yield by 43 kg/ha. Based on the soil particle size distribution measured for the six sampled locations (section 4.3.4.1), the available water capacity of the 0 - 30 cm soil depth was only about 34.1 mm to 45.3 mm. Theoretically, it would only require approximately 5 mm per day of consecutive rain for the water to leave the root zone to depth in some coarse-textured soils. To improve nitrate uptake by crops, it is generally recommended to split fertiliser applications into two or more applications and apply only a small amount of slow-release N fertiliser at each planting.

The accumulated rainfall in the seventeenth week after planting at the reproductive stage had a small negative influence on yield (Figure 4-15b). The estimated coefficient of 0.003 ($p<0.05$) in the linear regression model suggests that each 1 mm increase in the rainfall within this period could decrease yield by 3 kg/ha. An excessive amount of rainfall can reduce maize yield through direct physical damage to the plants and other processes associated with those soils which have poor drainage (waterlogging, ponding, overland flow, and prevention of optimal harvesting events) that are detrimental to the crop yield (Y. Li et al., 2019). Prolonged rainfall may also encourage diseases such as stalk rots and can result in lodging (the bending over of the crop stems near the ground level), which causes difficulties at harvest and increases yield losses.

The result in this thesis is different from a parallel study conducted in New Zealand by Holmes (2019) using farm-scale yield data recorded over 46 years by a local farmer in the Waikato Region, approximately 11km from the NCRS site. Holmes (2019) aimed to model maize yield based on the N applied (base, starter and side-dress applications), seasonal growing degree days [GDD] (spring and summer) and seasonal rainfall (spring and summer) using multiple linear regression with interaction terms. After comparing different models, his best model explains 69% of the yield variation and reported that the yield increases by 10 kg/ha for each additional mm of the rainfall in summer, whereas in this research, yield decreased by 3 kg/ha for each 1 mm increase in the rainfall in the seventeenth week after planting. The difference in results regarding the effect of summer rainfall on yield may have been caused by the lack of exact planting dates for comparing growth periods in his model. Holmes (2019)'s model is only explanatory at the current stage and cross-validation is still required to evaluate how well this model would perform for predicting actual yield. Also, no data filtering appears to have been applied to eliminate outliers.

This finding in this thesis on the effect of summer rain on maize yield is incomparable to J. Liu et al. (2001)'s study at the University of Illinois. This study identified the best combinations of factors that could predict maximum yield and suggests that yield was most sensitive to late July rainfall (mid-summer in the US) and nitrogen fertiliser (0 – 336 kg/ha). Despite the high temporal resolution of yield in J. Liu et al. (2001)'s

study (> 30 years continuous maize yield), monthly rainfall may be too coarse for modelling yield in New Zealand because the variable distribution of rainfall within a given month can lead to soil moisture deficits. As previously estimated (section 5.2.2.2.2), soil can only meet the plant water demand for another 10-13 days in the non-irrigated field examined in this thesis. Also, if there is no rain for three to eight hours after applying urea fertiliser, less urea is dissolved into the soil profile to be absorbed by the crop roots, and 10% of the total N applied may be lost via volatilisation as ammonia (NH_3) gas (Bishop & Manning, 2011).

In a non-irrigated field, yield variation is largely associated with the temporal pattern of rainfall to the crop at different growing stages, in conjunction with inherent soil texture variations within-field. To predict spatial yield, the temporal water distribution within-field, at different times, should be measured and incorporated into the prediction model. This temporal soil moisture information may potentially be derived from the Synthetic Aperture Radar (SAR) sensor data of Sentinel-1, which is calibrated by soil moisture probes installed in different soil zones. Because the SAR sensing is based on the backscattering of radar waves, it is not sensitive to cloud cover, solar illumination and atmospheric conditions (Belenguer-Plomer et al., 2019), allowing for year-round calibration and comparison. Given a fine spatial resolution (down to 5 m) and a 12-day revisit cycle, SAR sensor data could potentially provide some detail regarding temporal soil moisture distribution across the field. This can then be used as predictors of yield.

5.2.3.2.2 Inclusion of soil temperature data

The accumulated GDD within different growth periods contributed to temporal yield variability. Despite a slight decrease of yield caused by the accumulated GDD in the second week after planting (V2), Figure 4-14a suggests that accumulated GDD in the second week after planting between 40 and 80 degree Celsius tended to produce more stable yield than when the GDD range is less than 40 degree Celsius. A wet and a cool soil in the second week after planting is detrimental for crop establishment. Maize has a base temperature (the minimum temperature a plant requires to assimilate carbon dioxide through photosynthesis) range of 8 – 10°C (FAR, 2005). Warmer soil temperatures in spring are favourable for seed

germination and crop emergence, which can help the crop achieve an early crop maturity and avoid cold stress. In New Zealand, maize is ideally sown in early to mid-October, once the soil temperature has increased to above 10°C (Booker, 2009). Cool conditions during planting will retard crop emergence and seedling health. In coarser-textured soils, night-time temperatures in spring can drop significantly due to less water in the topsoil to regulate heat, even following warm days, inflicting extra stress on maize emergence (Alessi & Power, 1971). In practice, few New Zealand farmers measure soil temperature to decide the suitability for planting but base their decision mostly on the date, and soil moisture condition after spring rains: i.e. whether the ground is sufficiently dry for the planter to operate and also on the availability of planting machinery. In the sixteenth week after planting in this study (around early- to mid-February), yield increases with an increase of GDD and suggests that every increase of 1 degree Celsius in the accumulated GDD during this period may cause an increase of 110 kg/ha in yield (Figure 4-14b), whereas in Holmes (2019)'s model, the yield increases by 8.2 kg/ha for each additional degree of Summer GDD. The large difference in results from these models may have been caused by different yield response on different textured soils.

5.2.3.2.3 Inclusion of canopy cover data

The results of the analysis of variable rank in terms of influence showed that accumulated solar radiation was an important input identified by the RF model (Figure 4-12). Figure 4-13**a** indicates that an increase of the accumulated solar radiation in the fifth week after planting by 1 kJ/m^2 tended to increase grain yield by 43 kg/ha, suggested by the estimated coefficient. Maize is a C-4 plant that is efficient at converting radiation into biomass (about 16 kg dry matter/ha for every 1 kJ of radiation intercepted (FAR, 2009). Brown et al., (2007) in New Zealand modelled maize yield for several regions (Ruakura, Whakatane, Lincoln and Gore) based on historical data and estimated that Whakatane, in the Bay of Plenty Region, produced the highest yielding potential compared to the other locations because of the relatively high annual average radiation and temperature.

Although the linear fit was poor, the modelled results (Figure 4-13b) suggested that maize-grain yield declined slightly by 13 kg/ha with an increase in accumulated radiation two weeks before tasselling (in the eighth week after planting) by 1 kJ/m^2, due to a less than fully developed canopy. The developmental stage of the crop canopy and leaf chlorophyll content controls the efficiency of radiation interception for photosynthesis. Strong sunlight can also have negative effects which increase evapotranspiration, increasing the risk of the crop experiencing moisture stress (Ferrante & Mariani, 2018). At the crop tasselling stage, the canopy is fully developed allowing maximum radiation interception to take place (Teixeira et al., 2010). Shading occurs at this stage limiting photosynthesis in the sub-canopy and nutrient transfer from stalks to the ear, which could result in 20% yield reduction due to fewer developed kernels produced per ear (Reed et al, 1988; W. Liu & Tollenaar, 2009). This thesis was not able to confirm the effect of canopy cover (and thus radiation) on yield due to a limited number of years' yield data.

5.3 General discussion

5.3.1 Understanding the importance of data filtering in precision farming

As previously discussed, (Chapter 5.2.1), filtering yield monitor data is an important process that can improve the quality of yield maps (i.e. how closely the interpolated maps represent the actual sampled values). To date, there has been an inadequate investigation of the impact of poor-quality data on the uptake of yield mapping practices. There is some evidence (Fountas et al., 2005; Griffin et al., 2008) to suggest that one reason that farmers do not undertake yield mapping is that they receive little assistance in handling the data, which involves cleaning to the data to remove errors before further steps, which allows more accurate information to be achieved. Griffin et al. (2008) provided evidence for this when they interviewed six North American farmers with detailed questions on how they would use yield monitor data. They found that yield maps with errors affected the farmers' confidence in undertaking subsequent management actions. This highlights the importance of reducing errors in yield data to inform farm management practices.

Software such as the Yield Editor has been developed by Sudduth and Drummond (2007) to mitigate sensing errors. However, upon examination of historical yield data from NCRS in this thesis using this software, it was clear that processing is difficult because of different commercial sensors, as well as the possibility of missing information such as timestamps (see Appendix 1 for an example data). Therefore, to address this issue, a small customised programme in R (Appendix 2) based on the principles developed by Sudduth et al. (2012) and by Spekken et al. (2013), was developed as part of this PhD study which may help to 'clean' historical yield monitor data automatically. The benefit of this programme is that only yield values and their coordinates are required and hence it can be applied to yield monitor data from differing crops. It has been used in other studies to filter machine harvest yield data for potatoes, which contain a large amount of local variation (Jiang et al., 2017).

5.3.2 Understanding the value of spatial data in precision farming

5.3.2.1 Soil and elevation mapping at the subfield scale

Soil EC deep (depth between 0 and 90 cm) was consistently ranked the most important predictor of maize yield in both the Random Forest and XGBoost models using data pooled from five fields over years (Figure 4-12). Figure 4-12a shows that in the period before V6 (six-leaf vegetative stage), the most important predictors in the Random Forest model were: "soilec_deep" (100%), followed by "soilec_shallow" (70%) and "elevation" (39%), in terms of their relative ranking. These were also important predictors in the growth periods after V6: "soilec_deep" (100%), "soilec_shallow" (79%) and "elevation" (61%) (Figure 4-12b). Proximal soil EC sensing is a relatively well-established technique in precision farming and many studies use this approach (Lund et al., 2000; Kitchen et al., 2003; Corwin et al., 2010). A key advantage of this technique, relative to traditional sampling methods such as grid sampling, is the collection of spatially dense sample points, which are beneficial for mapping at finer subfield scale for precision farming operations such as delineation of MZs for fertiliser application, detection of soil compaction (Payne, 2008), irrigation scheduling (Hedley & Yule, 2009), and determination of yield potential (Lund et al., 1999).

In New Zealand, soil survey and soil type mapping are typically conducted at the regional scale, and it is uncommon to find soil maps at the farm scale, let alone field scale (Lilburne et al., 2012). To delineate dynamic MZs using the purposed method in this thesis, soil texture needs to be measured at the subfield level. Although only three soil cores taken from selected LY and HY zones derived from historical yield data would not be sufficient to map spatial soil texture variations at the field scale (as discussed in section 5.2.2.2.1), soil texture measurements were correlated to soil EC sensor data (R^2 = 0.47 with the average silt and R^2 = 0.69 with the average clay fractions in the topsoil of 5 – 30 cm, respectively) and the subsequent impact of soil texture on crop yield in relation to temporal rainfall pattern was confirmed by conducting the predictor importance analyses. Undertaking soil texture analysis has also illustrated that there is potentially significant spatial variation in texture across a short spatial distance (38 m and 41 m), even within a 10-ha non-irrigated field. These results provide a rationale to use soil EC data as a predictor for spatial yield.

5.3.3 Applicability of subfield yield prediction models for precision farming

To determine the feasibility of applying the model to other similar fields that have no or limited data, prediction models were built using data pooled from all five study fields. The predictions were evaluated by leaving out one set of site-year data each term and using the remaining pooled data for training the models. The results (Table 4-11) showed that the Random Forest and XGBoost models developed using pooled data produced multiyear average R^2 values ranged from 0.04 to 0.47. These R^2 values were lower (by 0.02 – 0.18) than when they were constructed using data collected from individual fields (multiyear average R^2 = 0.06 – 0.50). Given these results of prediction, these trained models are likely to be field-specific and are less capable of predicting yield for other fields, thus emphasising to individual growers the need for collecting multi-year spatially variable yield data.

5.4 Research contributions to precision farming

From the above discussion, this thesis contributes to precision maize cropping in New Zealand by identifying the following three key points:

1. To improve the quality of yield maps and subsequent modelling analyses, erroneous data points that are caused by systemic or operational mistakes must be removed.

Software programmes such as Yield Editor have been developed and are widely used for data filtering. However, upon examination of historical yield monitor data in New Zealand, it was found that information such as GPS timestamps were missing, which made it difficult to apply existing filtering software. For this reason, some of the work in this thesis involved the creation of a customised automatic filtering algorithm in R and validated its performance using geostatistical variogram analyses and 10-fold cross-validation. The results suggest that the filtering programme was effective at improving the accuracy of spatial interpolation for yield mapping, with only the coordinates and yield supplied as inputs. Therefore, this programme can be applied to other spatial data in precision farming with different structures, providing the opportunity for greater use of data. However, it should be noted that data filtering will not remediate data from an inadequately calibrated sensor at the start of each harvest, which still presents challenges to the further uptake of precision farming.

2. By examining the spatial and temporal variability of spatial yield maps, the delineated crop management zones using spatial data and conducted zonal soil sampling can be used to determine the applicability of spatial data for precision farming in New Zealand.

To understand the cause of spatial yield variability, soil texture variations within-field were investigated with samples taken from each zone. The effect of spatial soil variability and field elevation on crops was discussed in relation to temporal rainfall to justify the use of soil EC, elevation and temporal weather data as predictors for spatial yield. Low and medium resolution satellite images (Landsat-8 and Sentinel-2) were also collected and delineated into zones to correlate with soil and yield zones. The results support the inclusion of field elevation and soil EC maps on smaller-scale arable fields (<10 ha), which could encourage

further uptake of delineating management zones and zonal soil sampling in New Zealand for commercial variable rate applications.

3. Examining the viability of predicting spatial yield at the subfield level can help to delineate dynamic management zones and inform crop management inputs.

Earlier studies often lacked focus on the spatiotemporal interactions between soil, weather variables and crop. In addition to the models tested in previous studies, this study compared the performance of different supervised machine learning algorithms. By computing the variable importance using data pooled from different maize fields in the Waikato, significant contributors to the models were identified for further improvement of the model prediction. The results demonstrated the viability of predicting spatial yield at the subfield level for delineating dynamic MZs, which provided an alternative to mid-season variable rate input prescription, using data that is readily available and inexpensive. With the increasing availability of sensor-derived data such as yield maps for additional years, staged canopy cover derived from aerial photographs and detailed soil nutrient maps derived from hyperspectral imagery, the accuracy of the yield prediction models should be improved over time and provide insight into managing spatial yield variability within-field.

Several new questions were raised within the framework of this research but only a few of these could be addressed. These additional questions need to be addressed in future research:

- What would be the filtering performance using the developed algorithm when compared to other well-established programmes such as Yield Editor if the information required is intact?
- Can this modelling approach be applied to other crops in New Zealand?
- Would the impacts of spatiotemporal interactions on spatial yield still be consistent with more years of yield and meteorological data?
- Given the spatial soil texture variations and seasonal and within-season rainfall patterns, what would be the environmental impacts of applying variable rate N fertiliser in maize production?

The case study outlined in this thesis is potentially relevant in some developed countries where small crop farms are key players in the industry. These farms collect a large amount of subfield data over time and are driven to minimise the costs of production using data-processing techniques. Overall, this study utilised existing sensor-derived data in New Zealand and demonstrated a novel application of machine learning models which highlighted the value of spatial data collection to assist in more efficient use of production inputs and sustainable arable farming systems.

5.5 Summary

This chapter extrapolates from current findings and contrasts with published literature. Spatiotemporal data collected from various sources and utilised to predict spatial yield at a subfield level, to allow the prescribing crop management inputs such as seed and fertiliser.

The findings of this study demonstrate that:

- The customised spatial filtering algorithms programmed in R proved effective at improving the accuracy of yield maps, as indicated by the reduction of prediction errors. The benefit of this programme is that only yield values and their coordinates are required and hence it can be applied to yield monitor data from differing crops.
- Examining spatial and temporal yield variability reaffirmed the effect of field elevation, soil EC and temporal rainfall on yield, justifying the use of these datasets as yield predictors. When yield maps are absent, there is a potential to use medium-resolution (10 m) satellite images as proxies to yield maps for delineating MZs. By delineating zones from satellite images, soil EC survey and soil sampling can be undertaken from each zone for soil texture analysis, which provides a basis for prescribing seeding and fertiliser recommendation.
- This research recognised the feasibility of predicting yield at a subfield level using cheaply accessible data and provides a statistical basis for delineating dynamic MZs. The predictor importance analyses suggest that potential improvements to the yield prediction models can be made by incorporating soil moisture, soil temperature and canopy cover data at critical stages of crop development. However, given these results of prediction, these trained models are likely to be field-specific and are less capable of predicting yield for other fields.

Chapter 6 Summary, implications and limitations

6.1 Introduction

The research question posed in Chapter 1 was addressed in the research chapters and revisited in the discussion through an evaluation of model and analysis performance.

This research hypothesized that:

The process of delineating site-specific management zones can be improved by modelling spatiotemporal interactions between spatial crop yield and other complementary factors.

The research objectives related to the following issues and gaps in knowledge identified around the use of yield monitor data at the small field scale for crop production in New Zealand i.e.:

- To develop a filtering algorithm to improve maize yield mapping precision.
- To identify appropriate spatiotemporal yield predictors by examining historical yield maps, delineating subfield management zones, and undertaking soil sampling.
- To determine the viability of predicting dynamic maize yield at the subfield spatial scale using supervised machine learning algorithms.

This chapter outlines the summary of conclusions and their wider implications and identification of thesis limitations using thesis results and limitation to provide directions for future research.

6.2 Summary of conclusions

Precision farming is a data-driven management philosophy about applying the right input, at the right amount, at the right place and the right time. While precision agriculture is not new to the New Zealand arable community, this thesis demonstrates how sensor-derived data captured within-field can be fully utilised to inform crop management decisions. The work presented particularly addressed the aspects of

precision agriculture that focus on “the right amount” of inputs and knowing when the “the right time” might be with the predictions of spatial yield made at the subfield level.

These are the conclusions drawn from Chapter 5:

- A methodology approach taken in this thesis was to provide a customised automatic filtering algorithm that could work without this information. Its effect on the accuracy of the interpolated maps was validated using geostatistical variogram analyses and 10-fold cross-validation. The results (for example, the nugget/sill ratios were reduced from 0.86-0.95 to 0.54-0.58; RMSE was reduced from 4.41-4.59 to 0.96-1.43 for the interpolated yield in 2014, and 6.61-7.27 to 1.34-1.43 for the interpolated yield in 2015) suggest that the filtering programme was effective at improving the accuracy of spatial interpolation for yield mapping, with only the spatial coordinates and yield as input parameters.
- To identify important spatiotemporal yield predictors, this research project examined the spatial and temporal yield variability from historical yield maps. To achieve this, historical yield maps were delineated into yield productivity zones categorised as relative high yielding (HY) and low yielding (LY). To verify the effect of soil texture on spatial maize yield over the years, zonal soil sampling (based on the yield zones) and soil texture analysis were undertaken. The soil EC measurements were highly correlated (R^2 = 0.69; profile average) to the clay content in the topsoil (5-30 cm). These findings support those studies by other researchers, which may have caused spatial yield variability under a non-irrigated system.
- To determine the relationship between soil EC and spatial yield, the soil EC maps were correlated to the spatial yield map of each year. The yield response to soil EC appears to be a quadratic curve, with the reasonable fit (R^2 = 0.44 – 0.92), with higher ECs tending to produce better yield due to greater soil water holding capacity and being less susceptible to plant moisture stress. The variation in yield response between years suggests that soil EC should be discussed in relation to

the seasonal rainfall data so that soil EC measurements can be used as a spatial yield predictor more effectively.

- A reasonable agreement (71%, κ = 0.42) were found between MZs derived from multispectral images (Landsat-8 (30 m) and Sentinel-2 (10 m) and yield productivity zones. This result demonstrates that optical satellite images can be used as proxies to yield maps for delineating MZs when yield maps are absent. These results were consistent with other studies using high-resolution RapidEye images to delineate MZs.
- This research recognised the feasibility of predicting yield at a subfield level using cheaply accessible data and provides a statistical basis for delineating dynamic MZs. Compared to other models tested such as a feedforward neural network, the tree-based algorithms such as random forest and XGBoost tended to produce better statistical predictions (R^2 = 0.08 - 0.50) because of their ability to model multiple spatiotemporal interactions. Meteorological factors such as early seasonal rainfall at the crop vegetative stage (in the fifth week and seventeenth week after planting), GDD (the second week and sixteen-week after planting) and solar radiation (the fifth and eighth week after planting) were identified as important predictors of spatial yield, based on the relative importance derived from the prediction models.

6.3 Research limitations

This research was limited in several aspects:

1. Only three soil cores were taken from each yield-based LY and HY zone for investigating the soil texture variations within the field. This sampling density would not sufficiently represent these production zones. Depending on funding constraint, future research may look at more detailed soil sampling for mapping soil texture variability, and then correlate the texture to soil EC and yield.

2. More related variables such as soil pH should be included to improve the model predictions of spatial yield within-field. There is a potential influence of pH variation, and therefore nutrient availability, on

maize yield production within NCRS. The soil test results (Appendix 4 and 5) indicate that, for the whole field, pH values ranged from 6.3 to 7.6, slightly higher than optimum, and will potentially pose a risk in terms of zinc, manganese, and phosphorus availability to the growing maize crop (Longhurst et al., 2018). Soil pH data derived from Veris MSP measurement has been used for mapping previously (Holmes and Jiang, 2017) and can be used for future research.

3. This study did not investigate the impact of shelterbelt trees on yield. Some hedgerow species such as pine trees are known to require a large amount of moisture in summer with their more extensive root systems, which may compete with the crop nearby. The crop adjacent to the shelterbelt may also experience sunlight suppression due to the potential shading effect if the tree canopy has not been pruned adequately. At NCRS, sunlight is blocked in the late afternoon. In future research, it would be useful to look in detail at how the spatial yield is influenced by the shelterbelt trees. Shading is one of the features that can be derived from satellite images as a predictor for yield.

4. Another factor that might influence spatial yield in the field boundaries is that the soils in the headlands may also receive twice the compaction as the centre of the field because heavy farm vehicles (such as planters and harvesters) exit, turn around and then re-enter the field in these locations. Generally, fine-textured soils when wet, with fewer pore spaces, are more susceptible to compaction than coarse-textured soils, which have more pore spaces. There may be a spatial pattern in the level of compaction around the headlands associated with soil texture variations.

5. Limited years (3 to 7 years) of yield data collected from maize fields in Waikato were used to build yield prediction models due to historically poor yield sensor calibration, which may not be sufficient to capture temporal variability of yield. Yield monitor data from additional years and rotation crops should be analysed to provide further validation to the findings of this research. With more years of yield monitor data and crop images uploaded into the system, the yield prediction model for the individual field can become more accurate over time.

6.4 Recommendations for future research

Although the implementors of precision farming will be arable farmers, the data analysis tasks are mostly undertaken by farm consultants, researchers with specialist skills, and possibly agricultural contractors. To make these prediction models more accessible by farmers and consultants, there is a need to develop machine learning-based decision-supporting tools specifically for precision farming; For example, the trained models in this research can be embedded in a web application such as R “shiny app” with a simple and intuitive user interface. After receiving the data from farmers, the R “shiny app” can potentially initiate data analyses automatically, such as data filtering, mapping, mining, and the creation of management zones and providing crop management input recommendations. With more years of yield monitor data and crop images uploaded into the system, the yield prediction model for the individual field should be able to self-calibrate and become more accurate over time. The application of ML on precision farming data is likely to continue in research in the future.

This thesis has provided a case study on the analyses of on-site historical yield data. There is also a need to keep encouraging farmers to calibrate their yield sensors for developing accurate yield prediction models. Once the yield data is fully utilised for optimising crop management inputs, its potential economic and environmental value should be recognised by the arable industry worldwide.

6.5 Wider implications

The findings of this research suggest the following opportunities:

- The ability to improve product quality as well as quantity

To demonstrate the profitability of precision farming management practices for arable production in New Zealand, research may focus on product quality instead of just quantity. For example, in maize-grain production, the profitability is almost completely defined by yield. However, for wheat grain, the protein content is an important parameter for milling. The protein content of wheat is often increased with the increase of applied N. However, over-application of N may result in too high protein levels, decreased yield, increased disease incidence as well as posing an environmental risk. Therefore, there is a need to find the optimum for applied N rates that achieve a good balance between the quantity (yield) and quality (protein content) of wheat. Having the ability to predict the spatial gross margins of wheat at a subfield level will give the advantage of being able to prescribe N-rate before several in-season applications are applied, without overuse or the crop having limited supply of N.

- The ability to reduce environmental impacts on vegetable crop production

This statistical modelling technique can be applied to optimise inputs used in the production of vegetable crops. For example, field crop potatoes generally have shallower rooting than arable crops, a large demand for N, and a need for frequent irrigation in many regions. A high amount of fertiliser (up to 230 kg N/ha) can be used, especially for winter-planted varieties that are grown for table consumption. There is considerable potential for nitrate leaching to occur from these crops due to winter rainfall, if excessive fertiliser is applied. Excessive N concentrations in the tubers may also decrease tuber dry matter percentage (DM%) and density. Therefore, having the ability to predict marketable tuber yield during the production phase gives the advantage of being able to optimise crop management inputs spatially and reduce the risk of N leaching.

The researcher presented a preliminary case study on tuber yield mapping at 7th Asian-Australasian Conference on Precision Agriculture in 2017 (Jiang et al., 2017). This thesis encourages further investigations into the possibility of predicting spatial tuber yield before the harvest, which could provide a yield prediction model customised to New Zealand production.

6.6 Concluding remark

This research aimed to determine if the process of delineating site-specific management zones in maize crops could be improved by modelling spatiotemporal interactions between spatial and other complementary factors affecting yield. The findings of this research demonstrated the viability of predicting spatial yield based on a variety of spatiotemporal predictors (such as soil EC, rainfall) associated with individual fields that are cheaply available for crop producers in New Zealand.

The growing development of sensor technologies continues to make collecting a massive amount of data cheaper and easier. The advancement of sensor technologies has improved field sampling efficiencies and reduced sampling and analysis costs. Accurate point clouds can be collected and applied to subfield measurement and analysis, and medium to high-resolution satellite multispectral imagery is readily available for in-season crop monitoring. Spatial information collected from various sources can be integrated into precision applications (planting, spraying, fertiliser application) with variable rate technologies for varying inputs within-field. The increase of computer processing powers and machine learning models provides the possibility to extract knowledge from layers of information. Based on these trends, this study extrapolated from the existing sensor technologies and applied machine learning predictions that demonstrate the potential to fulfil the philosophy of precision farming and lead to more automated and sustainable farming systems in the future.

References

Abendroth, L., Elmore, R., Boyer, M., & Marlay, S. (2011). Corn growth and development.

Adams, M., Philpot, W., & Norvell, W. (1999). Yellowness index: an application of spectral second derivatives to estimate chlorosis of leaves in stressed vegetation. *International Journal of Remote Sensing*, 20(18), 3663-3675.

Alessi, J., & Power, J. (1971). Corn emergence in relation to soil temperature and seeding depth 1. *Agronomy Journal*, 63(5), 717-719.

Basso, B., Cammarano, D., Fiorentino, C., & Ritchie, J. (2013). Wheat yield response to spatially variable nitrogen fertilizer in Mediterranean environment. *European journal of agronomy*, 51, 65-70.

Basso, B., Dumont, B., Cammarano, D., Pezzuolo, A., Marinello, F., & Sartori, L. (2016). Environmental and economic benefits of variable rate nitrogen fertilization in a nitrate vulnerable zone. *Science of the total environment*, 545, 227-235.

Bausch, W., & Diker, K. (2001). Innovative remote sensing techniques to increase nitrogen use efficiency of corn. *Communications in soil science and plant analysis*, 32(7-8), 1371-1390.

Bausch, W., & Duke, H. (1996). Remote sensing of plant nitrogen status in corn. *Transactions of the ASAE*, 39(5), 1869-1875.

Beef + Lamb New Zealand (2019). Fact sheet. Retrieved from https://beeflambnz.com/sites/default/files/factsheets/pdfs/fact-sheet-128-managing-winter-crops-during-grazing.pdf (accessed January 2020)

Belenguer-Plomer, M., Tanase, M., Fernandez-Carrillo, A., & Chuvieco, E. (2019). Burned area detection and mapping using Sentinel-1 backscatter coefficient and thermal anomalies. *Remote Sensing of Environment*, 233, 111345.

Bezdek, J. (1973). Cluster validity with fuzzy sets. Retrieved from https://www.tandfonline.com/doi/pdf/10.1080/01969727308546047 (accessed February 2018)

Birrell, S., Sudduth, K., & Borgelt, S. (1996). Comparison of sensors and techniques for crop yield mapping. *Computers and Electronics in Agriculture*, 14(2-3), 215-233.

Bishop, P., & Manning, M. (2010). Urea volatilisation: the risk management and mitigation strategies. Palmerston North, New Zealand: Fertilizer and Lime Research Centre, Massey University.

Blackmer, T., & Schepers, J. (1995). Use of a chlorophyll meter to monitor nitrogen status and schedule fertigation for corn. *Journal of production agriculture*, 8(1), 56-60.

Blackmore, S. (2000). The interpretation of trends from multiple yield maps. *Computers and electronics in agriculture*, 26(1), 37-51.

Blackmore, S., & Marshall, J. (1996, January). Yield mapping; errors and algorithms. In *Proceedings of the Third International Conference on Precision Agriculture* (pp. 403-415). Madison, WI, USA: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.

Blackmore, S., Godwin, R., & Fountas, S. (2003). The analysis of spatial and temporal trends in yield map data over six years. *Biosystems Engineering*, 84(4), 455-466.

Blackmore, S., Stout, B., Wang, M., & Runov, B. (2005). Robotic agriculture–the future of agricultural mechanisation. Paper presented at the Proceedings of the 5th European conference on precision agriculture.

Blasch, G., & Taylor, J. (2018). Multi-temporal Yield Pattern Analysis–Adaption of Pattern Recognition to Agronomic Data.

Blasch, G., Spengler, D., Itzerott, S., & Wessolek, G. (2015). Organic matter modeling at the landscape scale based on multitemporal soil pattern analysis using RapidEye data. *Remote Sensing*, 7(9), 11125-11150.

Bongiovanni, R., & Lowenberg-DeBoer, J. (2004). Precision agriculture and sustainability. Precision agriculture, 5(4), 359-387.

Booker, J. (2009). Production, distribution and utilisation of maize in New Zealand: a dissertation submitted in partial fulfilment of the requirements for the degree of Masters [ie Master] of Applied Science at Lincoln University (Doctoral dissertation, Lincoln University).

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

Breiman, L. (2017). Classification and regression trees: Routledge.

Brown, H., Li, F., Wilson, D., & Fletcher, A. (2007). Geographical and seasonal variation in potential forage production in New Zealand. In: Meeting the Challenges for Pasture-Based Dairying. Proceedings of the Australasian Dairy Science Symposium, September 2007. 343-349

Burgess, T., & Webster, R. (1980). Optimal interpolation and isarithmic mapping of soil properties: i the semi-variogram and punctual kriging. *Journal of soil science*, 31(2), 315-331.

Buschmann, C., & Nagel, E. (1993). In vivo spectroscopy and internal optics of leaves as basis for remote sensing of vegetation. *International Journal of Remote Sensing*, 14(4), 711-722.

Butts-Wilmsmeyer, C., Seebauer, J., Singleton, L., & Below, F. (2019). Weather during key growth stages explains grain quality and yield of maize. *Agronomy*, 9(1), 16.

Buytaert, W., Celleri, R., Willems, P., De Bievre, B., & Wyseure, G. (2006). Spatial and temporal rainfall variability in mountainous areas: A case study from the south Ecuadorian Andes. *Journal of hydrology*, 329(3-4), 413-421.

Cakir, R. (2004). Effect of water stress at different development stages on vegetative and reproductive growth of corn. *Field Crops Research*, 89(1), 1-16.

Cannon, R., Dave, J., & Bezdek, J. (1986). Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE transactions on pattern analysis and machine intelligence*, (2), 248-255.

Chan, S., & Treleaven, P. (2015). Continuous model selection for large-scale recommender systems. In *Handbook of Statistics* (Vol. 33, pp. 107-124). Elsevier.

Chappell, P. (2013). The climate and weather of Waikato. NIWA Science and Technology Series 61, 40.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1-4.

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

Claydon, J. (1989). Determination of particle size distribution in fine-grained soil: pipette method. Division of Land and Soil Sciences Technical Record (New Zealand).

CliFlo: NIWA's National Climate Database on the Web. http://cliflo.niwa.co.nz/. Retrieved 26-May-2019

Colvin, T., & Arslan, S. (2000). A review of yield reconstruction and sources of errors in yield maps. *Paper presented at the Proceedings of the 5th International Conference on Precision Agriculture*, Bloomington, Minnesota, USA, 16-19 July 2000.

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

Corwin, D., & Lesch, S. (2003). Application of soil electrical conductivity to precision agriculture: theory, principles, and guidelines. *Agronomy Journal*, 95(3), 455-471.

Corwin, D., Lesch, S., Segal, E., Skaggs, T., & Bradford, S. (2010). Comparison of sampling strategies for characterizing spatial variability with apparent soil electrical conductivity directed soil sampling. *Journal of Environmental & Engineering Geophysics*, 15(3), 147-162.

Cushnahan, M., Wood, B., & Yule, I. (2017, October). Is big data driving a paradigm shift in precision agriculture?. In *Proceedings of the 7th Asian-Australasian Conference on Precision Agriculture*, Hamilton, New Zealand (pp. 16-18).

Dash, J., Watt, M., Pearse, G., Heaphy, M., & Dungey, H. (2017). Assessing very high-resolution UAV imagery for monitoring forest health during a simulated disease outbreak. *ISPRS Journal of Photogrammetry and Remote Sensing*, 131, 1-14.

De Benedetto, D., Castrignano, A., Diacono, M., Rinaldi, M., Ruggieri, S., & Tamborrino, R. (2013). Field partition by proximal and remote sensing data fusion. *Biosystems engineering*, 114(4), 372-383.

Demmel, M. (2013). Site-specific recording of yields Precision in Crop Farming (pp. 313-329): Springer.

Denmead, O., & Shaw, R. (1960). The Effects of Soil Moisture Stress at Different Stages of Growth on the Development and Yield of Corn 1. *Agronomy Journal*, 52(5), 272-274.

Dierke, C., & Werban, U. (2013). Relationships between gamma-ray data and soil properties at an agricultural test site. *Geoderma*, *199*, 90-98.

Doerge, T. (1999). Defining management zones for precision farming. *Crop Insight*. Vol 8, No. 21. Pioneer Hybrids.

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

Dubayah, R., & Drake, J. (2000). Lidar remote sensing for forestry. *Journal of Forestry*, *98*(6), 44-46.

Eberhardt, I., Schultz, B., Rizzi, R., Sanches, I., Formaggio, A., Atzberger, C., & José Barreto Luiz, A. (2016). Cloud cover assessment for operational crop monitoring systems in tropical areas. *Remote Sensing*, 8(3), 219.

Ekanayake, D., Owens, J., Holmes, A., & Werner, A. (2018). Delineation of 'Management Classes' within Non-Irrigated Maize Fields Using Readily Available Reflectance Data and Their Correspondence to Spatial

Yield Variation. Paper presented at the *14th International Conference on Precision Agriculture*, Montreal, Quebec, Canada.

Ferrante, A., & Mariani, L. (2018). Agronomic management for enhancing plant tolerance to abiotic stresses: High and low values of temperature, light intensity, and relative humidity. *Horticulturae*, 4(3), 21.

Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489*, 237-246.

Fleming, K., Heermann, D., & Westfall, D. (2004). Evaluating soil color with farmer input and apparent soil electrical conductivity for management zone delineation. *Agronomy Journal*, 96(6), 1581-1587.

Food and Agriculture Organization of the United Nations [FAO] (2015). Handbook on remote sensing for agricultural statistics. Retrieved from http://www.fao.org/3/ca6394en/ca6394en.pdf.

Foundation for Arable Research (2005). GDDs and solar radiation changes over the last 50 years. Retrieved from https://www.far.org.nz/assets/files/uploads/35Mz_GDD_and_SR.pdf (accessed June 2018).

Foundation for Arable Research (2009). Best management practices for growing maize on dairy farms. Retrieved from https://www.waikatoregion.govt.nz/assets/PageFiles/19416/publications/FAR%20best%20management%20practices%20-%20web.pdf (accessed December 2019).

Foundation for Arable Research (2010). FAR Focus 4 - Irrigation Management for Cropping: A growers guide. Retrieved from https://www.far.org.nz/articles/282/far-focus-4-irrigation-management-for-cropping-a-growers-guide (accessed December 2019).

Fountas, S., Blackmore, S., Ess, D., Hawkins, S., Blumhoff, G., Lowenberg-Deboer, J., & Sorensen, C. G. (2005). Farmer experience with precision agriculture in Denmark and the US Eastern Corn Belt. *Precision Agriculture*, 6(2), 121-141.

Fraisse, C., Sudduth, K., & Kitchen, N. (2001). Delineation of site-specific management zones by unsupervised classification of topographic attributes and soil electrical conductivity. *Transactions of the ASAE*, 44(1), 155.

Fridgen, J., Kitchen, N., Sudduth, K., Drummond, S., Wiebold, W., & Fraisse, C. (2004). Management Zone Analyst (MZA) Software for Subfield Management Zone Delineation. *Agronomy Journal*, 96(1), 100-108.

Garson, G. (1991). A comparison of neural network and expert systems algorithms with common multivariate procedures for analysis of social science data. *Social Science Computer Review*, 9(3), 399-434.

Gausman, H., Rodriguez, R., & Richardson, A. (1976). Infinite Reflectance of Dead Compared with Live Vegetation 1. *Agronomy Journal*, 68(2), 295-296.

Geary, P. (2003). Evaluation of field surface topography for improved precision in farming. PhD Thesis, unpublished. Cranfield University at Silsoe, Silsoe, Bedford, UK.

Genetic Technologies Limited (2016). Water management of maize. Retrieved from https://www.pioneer.co.nz/maize-silage/product-information/silage-technical-insights/water-management-of-maize.html (accessed December 2019).

Genetic Technologies Limited (2019). Maize for grain economics. Retrieved from https://www.pioneer.co.nz/maize-grain/tools/maize-grain-economics/ (accessed 31 March 2019).

Georgi, C., Spengler, D., Itzerott, S., & Kleinschmit, B. (2018). Automatic delineation algorithm for site-specific management zones based on satellite remote sensing data. *Precision agriculture*, 19(4), 684-707.

Godwin, R., Richards, T., Wood, G., Welsh, J., & Knight, S. (2003). An economic analysis of the potential for precision farming in UK cereal production. *Biosystems Engineering*, 84(4), 533-545.

Greenwell, B., Boehmke, B., & McCarthy, A. (2018). A simple and effective model-based variable importance measure. arXiv preprint arXiv:1805.04755.

Griffin, T., Dobbins, C., & Lowenberg-DeBoer, J. (2007). Case study of on-farm trials, spatial analysis and farm management decision making. *Precision agriculture*, 7, 745-752.

Griffin, T., Dobbins, C., Vyn, T., Florax, R., & Lowenberg-DeBoer, J. (2008). Spatial analysis of yield monitor data: case studies of on-farm trials and farm management decision making. *Precision Agriculture*, 9(5), 269-283.

Grisso, R., Alley, M., Holshouser, D., & Thomason, W. (2005). Precision farming tools. Soil electrical conductivity. Retrieved from https://vtechworks.lib.vt.edu/handle/10919/51377.

Guastaferro, F., Castrignanò, A., De Benedetto, D., Sollitto, D., Troccoli, A., & Cafarelli, B. (2010). A comparison of different algorithms for the delineation of management zones. *Precision agriculture*, 11(6), 600-620.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*: Springer Science & Business Media.

Haynes, R., & Knight, T. (1989). Comparison of soil chemical properties, enzyme activities, levels of biomass N and aggregate stability in the soil profile under conventional and no-tillage in Canterbury, New Zealand. *Soil and Tillage Research*, 14(3), 197-208.

Heaton, J. (2008). Introduction to neural networks with Java. *Heaton Research*, Inc..(p.159)

Hedley, C., & Yule, I. (2009). Soil water status mapping and two variable-rate irrigation scenarios. *Precision Agriculture*, 10(4), 342-355.

Hedley, C., Yule, I., Tuohy, M., & Kusumo, B. (2010). Proximal sensing methods for mapping soil water status in an irrigated maize field Proximal Soil Sensing (pp. 375-385): Springer.

Heege, H. (2013). Site-Specific Soil Cultivation. In *Precision in Crop Farming* (pp. 143-170). Springer, Dordrecht.

Hewitt, A. (2004). Soil properties for plant growth. Landcare Research Science Series, 26.

Hiemstra, P., Pebesma, E., Twenhöfel, C., & Heuvelink, G. (2009). Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Computers & Geosciences, 35*(8), 1711-1721.

Holmes, A. (2019). 46 years of maize grain data – what can it tell us?. Unpublished report.

Holmes, A., & Jiang, G. (2017). Effect of variable rate lime application on autumn-sown barley performance. *Agronomy New Zealand*, 47, 37-45.

Holmes, A., & Jiang, G. (2018, June). Increasing profitability & sustainability of maize using site-specific crop management in New Zealand. In *Proceedings of the 14th International Conference on Precision Agriculture*, Montreal, Quebec, Canada.

Hörbe, T. d. A., Amado, T., Ferreira, A. d. O., & Alba, P. (2013). Optimization of corn plant population according to management zones in Southern Brazil. *Precision agriculture*, 14(4), 450-465.

Huete, A., Artiola, J., & Pepper, I. (2004). Environmental monitoring with remote sensing. Environmental Monitoring and Characterization, 11, 183.

Hurst, C., Lovell, S., Lund, T., & Holmes, A. (2015). Precise surveying of soil productivity indicators using on-the-go soil sensors. Moving Farm Systems to Improved Attenuation; Currie, LD, Burkitt, LL, Eds.

Jackson, R., & Huete, A. (1991). Interpreting vegetation indices. *Preventive veterinary medicine*, 11(3-4), 185-200.

Jeschke, M., Carter, P., Bax, P., & Schon, R. (2015). Putting variable-rate seeding to work on your farm. *Crop Insights*, 25, 1-4. Retrieved from pioneer.com (accessed December 2018)

Jiang, G., Yule, I., Grafton, M., & Holmes, A. (2017). How can we demonstrate the economic value of precision agriculture (PA) practices to New Zealand agriculture service providers and arable farmers?. Poster presented in the *7th Asian-Australasian Conference on Precision Agriculture*.

Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, 143, 23-37.

Kaul, T., & Grafton, M. (2017). Geostatistical Determination of Soil Noise and Soil Phosphorus Spatial Variability. *Agriculture*, 7(10), 83.

Kerry, R., & Oliver, M. (2003). Variograms of ancillary data to aid sampling for soil surveys. Precision *Agriculture*, 4(3), 261-278.

Khosla, R., Westfall, D., Reich, R., Mahal, J., & Gangloff, W. (2010). Spatial variation and site-specific management zones. In Geostatistical applications for precision agriculture (pp. 195-219). Springer, Dordrecht.

King, A., & Eckersley, R. (2019). Statistics for Biomedical Engineers and Scientists: How to Visualize and Analyze Data. Academic Press (pp. 1-21).

Kitchen, N., Drummond, S., Lund, E., Sudduth, K., & Buchleiter, G. (2003). Soil electrical conductivity and topography related to yield for three contrasting soil–crop systems. *Agronomy Journal*, 95(3), 483-495.

Kitchen, N., Sudduth, K., Myers, D., Drummond, S., & Hong, S. (2005). Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. *Computers and Electronics in Agriculture*, 46(1-3), 285-308.

Koch, B., Khosla, R., Frasier, W., Westfall, D., & Inman, D. (2004). Economic feasibility of variable-rate nitrogen application utilizing site-specific management zones. *Agronomy Journal*, 96(6), 1572-1580.

Kormann, G., Demmel, M., & Auerohammer, H. (1998). Testing stand for yield measurement systems in combine harvesters. Presented at July 12-16, 1988 and 1998 ASAE Annual International Meeting. Paper No. X. ASAE, 2950 Niles Rd., St. Joseph, MI 49085-9659 USA.

Krige, D. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6), 119-139.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, 28(5), 1-26.

Kuhn, M., Weston, S., Keefer, C., & Coulter, N. (2012). Cubist models for regression. *R package Vignette R package version 0.0, 18*.

Lal, R., & Stewart, B. (Eds.). (2015). Soil-specific farming: precision agriculture (Vol. 22). CRC Press. (p.4)

Landis, J., & Koch, G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Lark, R., & Stafford, J. (1997). Classification as a first step in the interpretation of temporal and spatial variation of crop yield. *Annals of Applied Biology*, 130(1), 111-121.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

Leroux, C., Jones, H., Taylor, J., Clenet, A., & Tisseyre, B. (2018). A zone-based approach for processing and interpreting variability in multi-temporal yield data sets. *Computers and Electronics in Agriculture*, 148, 299-308.

Li, J., & Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, 53, 173-189.

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., & Peng, B. (2019). Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. *Global change biology*, *25*(7), 2325-2337.

Liaw, A., & Wiener, M. (2018). Classification and regression based on a forest of trees using random inputs. *R Package*.

Licht, M., Lenssen, A., & Elmore, R. (2017). Corn (Zea mays L.) seeding rate optimization in Iowa, USA. *Precision Agriculture*, 18(4), 452-469.

Ließ, M., Glaser, B., & Huwe, B. (2012). Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. *Geoderma*, 170, 70-79.

Lilburne L., Hewitt A. & Webb T. (2012). Soil and informatics science combine to develop S-map: a new generation soil information system for New Zealand. *Geoderma*, 170: 232-238, 10.1016/j.geoderma.2011.11.012.

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

Liu, W., & Tollenaar, M. (2009). Physiological mechanisms underlying heterosis for shade tolerance in maize. *Crop Science*, *49*(5), 1817-1826.

Longhurst, B., Taylor, M., Williams, I. (2018). Long-term maize grain growing in Waikato – Factors affecting sustainability. In: *Farm environmental planning – Science, policy and practice*. (Eds L. D. Currie and C. L. Christensen). http://flrc.massey.ac.nz/publications.html. Occasional Report No. 31. Fertilizer and Lime Research Centre, Massey University, Palmerston North, New Zealand. 12 pages.

Lowe, D. (2010). Introduction to the landscapes and soils of the Hamilton Basin. In D.J. Lowe, V.E. Neall, M. Hedley, B. Clothier & A. Mackay (Eds.), Guidebook for pre-conference North Island New Zealand "Volcanoes to ocean" 26th-30th July 2010, 19th World Congress of Soil Science: soil solutions for a changing world: Brisbane Australia 1-6 August 2010. (pp. 1.14-1.61). Palmerston North, New Zealand: New Zealand Society of Soil Science

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

Luck, J., & Fulton, J. (2014). Precision agriculture: Best management practices for collecting accurate yield data and avoiding errors during harvest. Ext. Publ. EC2004. Univ. of Nebraska, Lincoln. http://extensionpublications.unl. edu/assets/pdf/ec2004. pdf (accessed 5 May 2018).

Luck, J., & Fulton, J. (2015). Improving yield map quality by reducing errors through yield data file post-processing. *Inst Agric Nat Resour*, *9*(10).

Lund, E., & Maxton, C. (2011, May). Proximal sensing of soil organic matter using the Veris® OpticMapper™. In *2nd Global Workshop on Proximal Soil Sensing*, Montreal, Quebec, Canada (pp. 15-19).

Lund, E., Christy, C., & Drummond, P. (1999). Practical applications of soil electrical conductivity mapping. *Precision agriculture*, 99, 771-779.

Lund, E., Christy, C., & Drummond, P. (2000). Using yield and soil electrical conductivity (EC) maps to derive crop production performance information. Paper presented at the *Proceedings of the 5th International Conference on Precision Agriculture*, Bloomington, Minnesota, USA, 16-19 July 2000.

Maldaner, L., Corrêdo, L., Tavares, T., Mendez, L., Duarte, C., & Molin, J. (2018, June). Identifying and filtering out outliers in spatial datasets. In *Proceedings of the 14th International Conference on Precision Agriculture, Montreal, QC, Canada* (pp. 24-27).

Malvić, T. (2009). Geostatistics as a group of methods for advanced mapping of geological variables in hydrocarbon reservoirs. *Annual 2009 of the Croatian Academy of Engineering*, *12*, 69-83.

Martínez-Casasnovas, J., & Arnó, J. (2018). Use of farmer knowledge in the delineation of potential management zones in precision agriculture: a case study in maize (Zea mays L.). *Agriculture*, 8(6), 84.

Mascagni, H., Boquet, D., & Bell, B. (2007). Influence of starter fertilizer on corn yield and plant development on Mississippi River alluvial soils. *Beter Crops*, *91*(2), 8-10.

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

McBratney, A., Whelan, B., & Shatar, T. (1997). Variability and uncertainty in spatial, temporal and spatiotemporal crop-yield and related data. In *Ciba Foundation Symposium* (pp. 141-160).

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

McHugh, M. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, *22*(3), 276-282.

McWilliams, D., Berglund, D., & Endres, G. (1999). Corn growth and management quick guide. Retrieved from https://library.ndsu.edu/ (accessed 5 May 2018).

Miller, P., Lubke, G., McArtor, D., & Bergeman, C. (2016). Finding structure in data using multivariate tree boosting. *Psychological methods*, 21(4), 583.

Millner, J., Roskruge, N., & Dymond, J. (2013). The New Zealand arable industry. *Ecosystem services in New Zealand: conditions and trends*, 102-114.

Minasny, B., & McBratney, A. (2002). FuzME version 3.0. Australian Centre for Precision Agriculture, The University of Sydney, Australia.

Ministry for Primary Industries (2012). Farm monitoring report 2012 – Horticulture monitoring: Canterbury arable cropping. Wellington, Ministry for Primary Industries. http://www.mpi.govt.nz/news-resources/publications (accessed December 2019).

Ministry for Primary Industry (2018). Situation and Outlook for Primary Industries (SOPI) June 2018. Retrieved from https://www.mpi.govt.nz/dmsdocument/29291/direct (accessed December 2019).

Molloy, L. (1998). Soils in the New Zealand Landscape: the Living Mantle. *New Zealand Soil Science Society*.

Moral, F., Terrón, J., & Da Silva, J. M. (2010). Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. *Soil and Tillage Research*, 106(2), 335-343.

Morgan, J., & Sonquist, J. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302), 415-434.

Muchow, R., Sinclair, T., & Bennett, J. (1990). Temperature and solar radiation effects on potential maize yield across locations. *Agronomy Journal*, 82(2), 338-343.

Mukaka, M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3), 69-71.

Murphy, D., Schnug, E., & Haneklaus, S. (1995). Yield mapping—A guide to improved techniques and strategies. Paper presented at the *Site-specific management for agricultural systems*.

Murphy, S. & Hård, J. (2017). Atmospheric correction of Sentinel 2 imagery in Google Earth Engine using Py6S. Retrieved from https://github.com/samsammurphy/gee-atmcorr-S2

Nichols, V., Ordonez, R., Wright, E., Castellano, M., Liebman, M., Hatfield, J., & Archontoulis, S. (2019). Maize root distributions strongly associated with water tables in Iowa, USA. *Plant and Soil*, *444*(1-2), 225-238.

Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win" every" machine learning competition? (Master's thesis, NTNU).

NIWA (2016). Overview of New Zealand's climate. Retrieved from https://niwa.co.nz/education-and-training/schools/resources/climate/overview (accessed January 2019).

Noi, P., Degener, J., & 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), 398.

Oliver, M. (2010). An overview of geostatistics and precision agriculture Geostatistical applications for precision agriculture (pp. 1-34): Springer.

Orimoloye, L., Sung, M., Ma, T., & Johnson, J. (2020). Comparing the effectiveness of deep feedforward neural networks and shallow architectures for predicting stock price indices. *Expert Systems with Applications*, 139, 112828.

Pantazi, X. , Moshou, D., Mouazen, A. M., Alexandridis, T., & Kuang, B. (2015, September). Data Fusion of Proximal Soil Sensing and Remote Crop Sensing for the Delineation of Management Zones in Arable Crop Precision Farming. In HAICTA (pp. 765-776).

Payne, J. (2008). Identification of Subsoil Compaction Using Electrical Conductivity and Spectral Data Across Varying Soil Moisture Regimes in Utah [All Graduate Theses and Dissertations. 26]. Retrieved from https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1025&context=etd.

Pinter Jr, P., Hatfield, J., Schepers, J., Barnes, E., Moran, M., Daughtry, C., & Upchurch, D. (2003). Remote sensing for crop management. *Photogrammetric Engineering & Remote Sensing*, 69(6), 647-664.

Ravensdown Limited (2019). Soil testing with ARL – price list. Retrieved from https://www.ravensdown.co.nz/media/4876/soil-testing-prices.pdf

Reed, A., Singletary, G., Schussler, J., Williamson, D., & Christy, A. (1988). Shading effects on dry matter and nitrogen partitioning, kernel number, and yield of maize. *Crop Science*, 28(5), 819-825.

Rhoades, J., Corwin, D., & Lesch, S. (1999). Geospatial measurements of soil electrical conductivity to assess soil salinity and diffuse salt loading from irrigation. *Geophysical Monograph-American Geophysical Union*, 108, 197-216.

Ripley, B., & Venables, W. (2016). nnet: Feed-forward neural networks and multinomial log-linear models. *R package version*, *7*.

Robert, P. (1993). Characterization of soil conditions at the field level for soil specific management. *Geoderma*, 60(1-4), 57-72.

Rodrigues Jr, F., Bramley, R., & Gobbett, D. (2015). Proximal soil sensing for precision agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in contrasting soils. *Geoderma*, 243, 183-195.

Rodriguez, J., Perez, A., & Lozano, J. (2009). Sensitivity analysis of k-fold cross-validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3), 569-575.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53-65.

Roy, D., Wulder, M., Loveland, T., Woodcock, C., Allen, R., Anderson, M., Kennedy, R. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote sensing of environment, 145*, 154-172.

Rumbal, P. (1978). Some effects of a variable water table on soil and pasture in Manawatu sand country. *New Zealand Journal of Experimental Agriculture*, 6(3), 241-249.

Rund, Q. (2018, February). 14th International Conference on Precision Agriculture. ISPA Newsletter, 52. Retrieved from https://ispag.org/site/newsletter/?id=52.

Saglio, P., Raymond, P., & Pradet, A. (1983). Oxygen transport and root respiration of maize seedlings: a quantitative approach using the correlation between ATP/ADP and the respiration rate controlled by oxygen tension. *Plant Physiology*, 72(4), 1035-1039.

Salinger, M. (1986). Nuclear winter: Impacts on the growing season in New Zealand. *Journal of the Royal Society of New Zealand*, 16(4), 319-333.

Saunders, C., & Saunders, J. (2012). The economic value of potential irrigation in Canterbury. Retrieved from http://researcharchive.lincoln.ac.nz/handle/10182/6973 (accessed December 2019).

Schimmelpfennig, D. (2016). Farm profits and adoption of precision agriculture (No. 1477-2016-121190). Retrieved from https://ageconsearch.umn.edu/ (accessed December 2019).

Schirrmann, M., Gebbers, R., Kramer, E., & Seidel, J. (2011). Soil pH mapping with an on-the-go sensor. *Sensors*, 11(1), 573-598.

Shanahan, J., Doerge, T., Snyder, C., Luchiari Jr, A., & Johnson, J. (2000). Feasibility of variable rate management of corn hybrids and seeding rates. Paper presented at the *Proceedings of the 5th International Conference on Precision Agriculture*, Bloomington, Minnesota, USA, 16-19 July 2000.

Shanahan, J., Schepers, J., Francis, D., Varvel, G., Wilhelm, W., Tringe, J., Major, D. (2001). Use of remote-sensing imagery to estimate corn grain yield. *Agronomy Journal*, 93(3), 583-589.

Simbahan, G., Dobermann, A., & Ping, J. (2004). Screening yield monitor data improves grain yield maps. *Agronomy Journal*, 96(4), 1091-1102.

Song, L., Jin, J., & He, J. (2019). Effects of severe water stress on maize growth processes in the field. *Sustainability*, *11*(18), 5086.

Spekken, M., Anselmi, A., & Molin, J. (2013). A simple method for filtering spatial data. *Precision agriculture'13* (pp. 259-266): Springer.

Stadler, A., Rudolph, S., Kupisch, M., Langensiepen, M., van der Kruk, J., & Ewert, F. (2015). Quantifying the effects of soil variability on crop growth using apparent soil electrical conductivity measurements. *European journal of agronomy*, 64, 8-20.

Statistics New Zealand (2018a). 2018 Census data. https://www.stats.govt.nz/tools/2018-census-place-summaries/waikato-region (accessed December 2018).

Statistics New Zealand (2018b). Agricultural production statistics June 2018. Retrieved from http:// www.stats.govt.nz (accessed December 2018).

Statistics New Zealand (2019). Agricultural production statistics June 2019. Retrieved from http:// www.stats.govt.nz (accessed December 2019).

Sudduth, K., & Drummond, S. (2007). Yield editor. *Agronomy Journal*, 99(6), 1471-1482.

Sudduth, K., Drummond, S., & Myers, D. (2012). Yield editor 2.0: Software for automated removal of yield map errors. Paper presented at the 2012 Dallas, Texas, July 29-August 1, 2012.

Sudduth, K., Drummond, S., Birrell, S., & Kitchen, N. (1996, January). Analysis of spatial factors influencing crop yield. In *Proceedings of the Third International Conference on Precision Agriculture* (pp. 129-139). Madison, WI, USA: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.

Sudduth, K., Kitchen, N., Bollero, G., Bullock, D., & Wiebold, W. (2003). Comparison of electromagnetic induction and direct sensing of soil electrical conductivity. *Agronomy Journal*, 95(3), 472-482.

Sudduth, K., Kitchen, N., Wiebold, W., Batchelor, W., Bollero, G., Bullock, D., & Thelen, K. (2005). Relating apparent electrical conductivity to soil properties across the north-central USA. *Computers and Electronics in Agriculture*, *46*(1-3), 263-283.

Taylor, J., Mason, M., Whelan, B., & McBratney, A. (2006). Determining optimum management zone-based seeding rates using on-farm experimentation and variable rate seeding technologies. Paper presented at the *USA International PA Conference*.

Taylor, J., McBratney, A., & Whelan, B. (2007). Establishing management classes for broadacre agricultural production. *Agronomy Journal*, 99(5), 1366-1376.

Teixeira, E., Brown, H., Chakwizira, E., & de Ruiter, J. (2010, November). Predicting yield and biomass nitrogen of forage crop rotations in New Zealand using the APSIM model. In *Proceedings of the 15th ASA Conference* (pp. 15-19).

Therneau, T., & Atkinson, E. (1997). An introduction to recursive partitioning using the RPART routines (Vol. 61, p. 452). Mayo Foundation: Technical report.

Thylén, L., & Murphy, D. (1996). The control of errors in momentary yield data from combine harvesters. *Journal of agricultural engineering research*, 64(4), 271-278.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), 234-240.

Tremblay, N. (2015). ISPA and precision agriculture around the world [PowerPoint slides]. Retrieved from http://past.infoag.org/abstract_papers/papers/paper_325.pdf.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of environment*, 8(2), 127-150.

Velandia, M., Buschermohle, M., Larson, J., Thompson, N., & Jernigan, B. (2013). The economics of automatic section control technology for planters: A case study of middle and west Tennessee farms. *Computers and electronics in agriculture*, 95, 1-10.

Walton, J. (2008). Subpixel urban land cover estimation. *Photogrammetric Engineering & Remote Sensing*, 74(10), 1213-1222.

Wang, X., Miao, Y., Dong, R., Chen, Z., Guan, Y., Yue, X., Mulla, D. (2019). Developing Active Canopy Sensor-Based Precision Nitrogen Management Strategies for Maize in Northeast China. *Sustainability*, 11(3), 706.

Webster, R., & Oliver, M. (1992). Sample adequately to estimate variograms of soil properties. *Journal of soil science*, 43(1), 177-192.

Whelan, B. (2011). A review of the history of Precision Agriculture in Australia and some future opportunities [PowerPointSlide]. Retrieved from http://docplayer.net/ (accessed December 2019). (p.3).

Whelan, B., & McBratney, A. (2003). Definition and interpretation of potential management zones in Australia. Paper presented at the Proceedings of the 11th Australian Agronomy Conference, Geelong, Victoria.

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

Wiegand, C., Richardson, A., Escobar, D., & Gerbermann, A. (1991). Vegetation indices in crop assessments. *Remote sensing of environment*, 35(2-3), 105-119.

Wilson, D, Muchow, R., & Murgatroyd, C. (1995). Model analysis of temperature and solar radiation limitations to maize potential productivity in a cool climate. *Field crops research*, *43*(1), 1-18.

Wilson, R. (2013). Py6S: A Python interface to the 6S radiative transfer model. *Computers And Geosciences*., 51(2), 166.

Yang, Y., Xu, W., Hou, P., Liu, G., Liu, W., Wang, Y., & Li, S. (2019). Improving maize grain yield by matching maize growth and solar radiation. *Scientific reports*, *9*(1), 1-11.

Zarco-Tejada, P. J., Hubbard, N., & Loudjani, P. (2014). Precision agriculture: An opportunity for EU farmers—potential support with the CAP 2014-2020. *Joint Research Centre (JRC) of the European Commission*.

Zhao, Y., & Cen, Y. (2013). *Data mining applications with R*: Academic Press.

Zhou, J., Li, E., Wei, H., Li, C., Qiao, Q., & Armaghani, D. J. (2019). Random forests and cubist algorithms for predicting shear strengths of rockfill materials. *Applied Sciences*, 9(8), 1621.

Appendix 1 Example of yield monitor data

Table A - 1 ***A subset of the yield monitor data (NCRS 2018)***

	coords.x1	coords.x2	Field	Dataset	Productivi	Obj__Id	Swth_Wdth_	Distance_m	Duration_s	Track_deg_	Elevation_	Area_Count	Time	Y_Offset_m	Pass_Num	Moisture__	Crop_Flw_M	Speed_km_h	Crop_Flw_V	Yld_Mass_W	Yld_Mass_D	Yld_Vol_We	Yld_Vol_Dr	Prod_ha_h_	Date
1	175.372415189524	-37.8369277327349	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	1	6.08	0.89	1	32	55.26	On	20180502	0	1	5.386	0.721	5.456	0	0.37	0.37	513.108	513.108	1.948	2018/05/02
2	175.372419686134	-37.8369206302505	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	2	6.08	0.87	1	30	55.24	On	20180502	0	1	17.172	2.056	5.675	0.001	1.08	1.059	1498.162	1468.518	1.904	2018/05/02
3	175.372423283422	-37.8369135277654	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	3	6.08	0.87	1	30	55.28	On	20180502	0	1	20.306	3.858	5.806	0.002	2.026	1.911	2810.421	2650.576	1.904	2018/05/02
4	175.372427420303	-37.8369061411802	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	4	6.08	0.89	1	28	55.36	On	20180502	0	1	21.052	5.659	5.771	0.002	2.905	2.714	4030.035	3765.245	1.948	2018/05/02
5	175.372430837727	-37.8368991807435	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	5	6.08	0.85	1	26	55.5	On	20180502	0	1	21.224	7.25	5.105	0.003	3.897	3.633	5406.121	5039.912	1.861	2018/05/02
6	175.372432816235	-37.836892220306	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	6	6.08	0.77	1	25	55.66	On	20180502	0	1	21.268	8.811	3.695	0.003	5.228	4.871	7252.787	6757.709	1.685	2018/05/02
7	175.372435694066	-37.836885259868	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	7	6.08	0.82	1	23	55.66	On	20180502	0	1	21.27	10.582	1.777	0.004	5.896	5.494	8179.614	7621.077	1.795	2018/05/02
8	175.372438751761	-37.8368774471306	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	8	6.08	0.91	1	22	55.58	On	20180502	0	1	21.27	12.714	0.75	0.005	6.383	5.947	8855.224	8250.553	1.992	2018/05/02
9	175.372442169185	-37.8368686400438	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	9	6.08	1.02	1	21	55.6	On	20180502	0	1	21.27	16.136	0.525	0.006	7.228	6.734	10026.88	9342.204	2.233	2018/05/02
10	175.372445586608	-37.8368586965575	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	10	6.08	1.15	1	20	55.64	On	20180502	0	1	21.27	20.729	0.415	0.008	8.235	7.673	11424.925	10644.785	2.517	2018/05/02
11	175.372448824168	-37.8368467643721	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	11	6.08	1.37	1	18	55.96	On	20180502	0	1	21.27	24.602	0.383	0.009	8.204	7.644	11381.93	10604.726	2.999	2018/05/02
12	175.372451881863	-37.8368334116862	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	12	6.08	1.51	1	15	56.36	On	20180502	0	1	21.27	27.049	0.727	0.01	8.184	7.625	11353.646	10578.374	3.305	2018/05/02
13	175.372454939557	-37.8368204851475	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	13	6.08	1.45	1	13	56.42	On	20180502	0	1	21.27	27.439	1.667	0.011	8.646	8.055	11994.046	11175.044	3.174	2018/05/02
14	175.372457637523	-37.8368074165567	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	14	6.08	1.48	1	12	56.62	On	20180502	0	1	21.27	25.878	2.372	0.01	7.989	7.443	11082.381	10325.631	3.239	2018/05/02
15	175.372460515354	-37.8367943479635	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	15	6.08	1.47	1	11	56.6	On	20180502	0	1	21.27	23.882	1.825	0.009	7.422	6.915	10296.992	9593.872	3.217	2018/05/02
16	175.372463573049	-37.8367811373181	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	16	6.08	1.48	1	11	56.6	On	20180502	0	1	21.27	23.776	0.776	0.009	7.34	6.839	10182.42	9487.123	3.239	2018/05/02
17	175.372466450879	-37.8367680687202	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	17	6.08	1.5	1	10	56.64	On	20180502	0	1	21.02	25.563	2.099	0.01	7.786	7.277	10801.421	10095.814	3.283	2018/05/02
18	175.372470048168	-37.83675414782	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	18	6.08	1.57	1	10	56.7	On	20180502	0	1	20.478	26.824	3.306	0.01	7.806	7.346	10828.852	10190.911	3.436	2018/05/02
19	175.372474364913	-37.8367403689672	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	19	6.08	1.58	1	11	56.72	On	20180502	0	1	20.332	25.758	3.441	0.01	7.448	7.022	10332.793	9741.928	3.458	2018/05/02
20	175.372478861523	-37.8367267321619	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	20	6.08	1.57	1	13	56.8	On	20180502	0	1	20.3	21.555	3.875	0.008	6.273	5.916	8701.864	8207.557	3.436	2018/05/02
21	175.372482998405	-37.8367153681556	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	21	6.08	1.32	1	15	56.74	On	20180502	0	1	20.3	14.665	4.722	0.006	5.076	4.787	7041.71	6641.708	2.889	2018/05/02
22	175.372486595693	-37.8367071292499	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	22	6.08	0.86	1	16	56.78	On	20180502	0	1	20.3	9.006	5.018	0.004	4.785	4.513	6637.588	6260.542	1.882	2018/05/02
23	175.372487854744	-37.8367038620973	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	23	4.86	0.31	1	17	56.7	On	20180502	0	1	20.3	6.95	5.23	0.003	12.814	12.086	17776.413	16766.628	0.542	2018/05/02
24	175.37249792715	-37.8366861058312	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	24	6.08	1.26	1	19	56.72	On	20180502	0	2	19.04	1.636	4.532	0.001	0.593	0.568	823.026	788.546	2.758	2018/05/02
25	175.372503502947	-37.8366744577182	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	25	6.08	1.39	1	20	56.86	On	20180502	0	2	19.04	4.008	4.558	0.002	1.317	1.262	1827.486	1750.926	3.042	2018/05/02
26	175.372511057252	-37.8366620993525	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	26	6.08	1.53	1	23	57.3	On	20180502	0	2	19.04	7.43	4.632	0.003	2.219	2.126	3078.019	2949.07	3.349	2018/05/02
27	175.372520050472	-37.8366494568842	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	27	6.08	1.61	1	27	57.7	On	20180502	0	2	19.04	10.943	4.657	0.004	3.105	2.975	4307.836	4127.366	3.524	2018/05/02
28	175.372529583286	-37.8366372405645	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	28	6.08	1.61	1	29	57.82	On	20180502	0	2	19.04	13.78	4.319	0.005	3.91	3.746	5424.682	5197.424	3.524	2018/05/02
29	175.372540734879	-37.8366260185946	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	29	6.08	1.59	1	32	57.48	On	20180502	0	2	19.04	15.821	3.856	0.006	4.546	4.356	6306.683	6042.474	3.48	2018/05/02
30	175.372551706608	-37.8366153648242	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	30	6.08	1.54	1	35	57.48	On	20180502	0	2	19.04	17.997	3	0.007	5.339	5.116	7407.233	7096.918	3.371	2018/05/02
31	175.372564297116	-37.8366065577061	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	31	6.08	1.51	1	40	57.34	On	20180502	0	2	19.04	21.315	1.201	0.008	6.449	6.179	8946.825	8572.011	3.305	2018/05/02
32	175.372577247353	-37.836597750587	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	32	6.08	1.5	1	44	57.38	On	20180502	0	2	19.04	25.142	2.14	0.01	7.658	7.337	10623.829	10178.76	3.283	2018/05/02
33	175.372590737184	-37.8365889434668	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	33	6.08	1.54	1	48	57.4	On	20180502	0	2	19.04	28.475	3.85	0.011	8.448	8.094	11719.366	11228.401	3.371	2018/05/02
34	175.37260566593	-37.8365812727484	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	34	6.08	1.55	1	51	57.32	On	20180502	0	2	18.906	31.417	4.336	0.012	9.26	8.887	12846.803	12328.978	3.393	2018/05/02
35	175.372620234946	-37.8365733179284	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	35	6.08	1.58	1	54	57.3	On	20180502	0	2	18.822	33.698	4.558	0.013	9.744	9.361	13518.135	12986.688	3.458	2018/05/02
36	175.372634444234	-37.8365647949061	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	36	6.08	1.59	1	57	57.18	On	20180502	0	2	18.81	35.905	4.065	0.014	10.317	9.913	14312.7	13752.049	3.48	2018/05/02
37	175.372650452167	-37.8365585446891	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	37	6.08	1.58	1	61	57.2	On	20180502	0	2	18.81	37.346	2.202	0.014	10.799	10.376	14981.345	14394.502	3.458	2018/05/02
38	175.372665021184	-37.8365518683204	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	38	6.08	1.49	1	62	57.16	On	20180502	0	2	18.81	36.776	0.605	0.014	11.276	10.835	15643.623	15030.837	3.261	2018/05/02
39	175.372678151285	-37.8365450499006	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	39	6.08	1.37	1	62	57.2	On	20180502	0	2	18.81	34.224	0.605	0.013	11.413	10.966	15833.313	15213.097	2.999	2018/05/02
40	175.372690741794	-37.8365389417324	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	40	6.08	1.38	1	62	57.18	On	20180502	0	2	18.81	31.357	0.605	0.012	10.381	9.975	14401.803	13837.662	3.021	2018/05/02
41	175.372703332302	-37.8365328335637	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	41	6.08	1.69	1	61	57.14	On	20180502	0	2	18.81	28.925	0.605	0.011	7.82	7.513	10848.071	10423.135	3.699	2018/05/02
42	175.372715203353	-37.8365254469403	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	42	6.08	1.69	1	60	57.28	On	20180502	0	2	18.81	27.094	0.605	0.01	7.324	7.038	10161.271	9763.237	3.699	2018/05/02
43	175.372726354946	-37.8365170659629	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	43	6.08	1.5	1	57	57.4	On	20180502	0	2	18.81	26.403	0.605	0.01	8.042	7.727	11156.606	10719.584	3.283	2018/05/02
44	175.372736067624	-37.8365092531865	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	44	6.08	1.22	1	53	57.4	On	20180502	0	2	18.81	27.214	0.605	0.011	10.191	9.792	14138.245	13584.427	2.67	2018/05/02
45	175.372744701115	-37.8365018665608	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	45	6.08	1.12	1	50	57.36	On	20180502	0	2	18.81	28.73	0.605	0.011	11.72	11.261	16258.536	15621.663	2.451	2018/05/02
46	175.372750996369	-37.8364966106921	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	46	6.08	0.67	1	48	57.34	On	20180502	0	2	18.81	34.974	0.605	0.014	23.849	22.915	33085.571	31789.557	1.466	2018/05/02
47	175.37274775881	-37.8364991676012	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	47	6.08	0.9	1	47	57.4	On	20180502	0	2	18.152	18.853	0.605	0.007	9.57	9.27	13277.144	12860.446	1.97	2018/05/02
48	175.372739844776	-37.8365055598738	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	48	3.65	1.03	1	48	57.46	On	20180502	0	2	17.73	12.053	0.605	0.005	8.906	8.671	12355.108	12029.05	1.353	2018/05/02
49	175.372403138609	-37.8369291532317	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	49	3.34	0.95	1	341	55.56	On	20180502	0	3	18.996	0.405	6.161	0	0.355	0.34	492.215	471.851	1.142	2018/05/02
50	175.372398641998	-37.8369204882008	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	50	4.56	1.05	1	341	55.58	On	20180502	0	3	19.148	1.426	6.943	0.001	0.827	0.792	1147.706	1098.158	1.724	2018/05/02
51	175.372393785659	-37.8369102606221	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	51	5.47	1.23	1	342	55.66	On	20180502	0	3	19.18	2.837	7.558	0.001	1.171	1.12	1624.914	1554.149	2.422	2018/05/02
52	175.37238892932	-37.8368993227932	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	52	6.08	1.28	1	343	55.68	On	20180502	0	3	19.18	5.164	7.847	0.002	1.843	1.763	2556.854	2445.503	2.802	2018/05/02
53	175.372384972304	-37.8368876747139	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	53	6.08	1.34	1	344	55.92	On	20180502	0	3	19.18	12.534	7.844	0.005	4.273	4.087	5928.423	5670.239	2.933	2018/05/02
54	175.372381914609	-37.8368767368817	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	54	6.08	1.25	1	346	56	On	20180502	0	3	19.18	20.849	7.976	0.008	7.62	7.289	10571.82	10111.414	2.736	2018/05/02
55	175.372379576371	-37.8368656569981	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	55	6.08	1.24	1	348	56	On	20180502	0	3	19.18	24.422	8.053	0.009	8.998	8.606	12483.127	11939.483	2.714	2018/05/02
56	175.372377957592	-37.8368542930131	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	56	6.08	1.28	1	352	56	On	20180502	0	3	19.18	23.912	8.098	0.009	8.535	8.163	11840.317	11324.668	2.802	2018/05/02
57	175.372377238134	-37.8368434972258	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	57	6.08	1.21	1	355	56.08	On	20180502	0	3	19.18	20.759	8.169	0.008	7.838	7.497	10874.125	10400.554	2.648	2018/05/02
58	175.372376878405	-37.8368368208828	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	58	6.08	0.49	1	357	56.12	On	20180502	0	3	19.18	16.406	8.307	0.006	15.297	14.631	21221.77	20297.556	1.073	2018/05/02
59	175.372379216643	-37.8368457700232	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	59	1.22	0.87	1	359	56	On	20180502	0	3	19.18	9.952	7.261	0.004	26.045	24.911	36132.295	34558.723	0.382	2018/05/02
60	175.372380655558	-37.8368244625443	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	60	6.08	1.35	1	3	56.1	On	20180502	0	3	18.792	19.619	6.357	0.008	6.639	6.381	9210.842	8852.001	2.955	2018/05/02
61	175.372382993795	-37.8368098314052	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	61	6.08	1.65	1	5	56.1	On	20180502	0	3	18.69	24.407	2.15	0.009	6.758	6.503	9375.493	9021.554	3.611	2018/05/02
62	175.372386411219	-37.836793211564	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	62	6.08	1.87	1	8	56.18	On	20180502	0	3	18.67	27.529	3.463	0.011	6.726	6.473	9330.722	8980.682	4.093	2018/05/02
63	175.372390188371	-37.8367754553193	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	63	6.08	2	1	10	56.28	On	20180502	0	3	18.67	29.751	4.857	0.012	6.796	6.541	9428.252	9074.553	4.378	2018/05/02
64	175.372394864846	-37.8367571308703	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	64	6.08	2.08	1	11	56.3	On	20180502	0	3	18.67	31.192	4.062	0.012	6.851	6.594	9504.729	9148.161	4.553	2018/05/02
65	175.37239954132	-37.8367386643667	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	65	6.08	2.11	1	12	56.3	On	20180502	0	3	18.67	33.158	1.194	0.013	7.18	6.91	9960.263	9586.605	4.618	2018/05/02
66	175.372404757388	-37.836719345558	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	66	6.08	2.19	1	12	56.36	On	20180502	0	3	18.946	35.605	0.126	0.014	7.428	7.125	10304.528	9884.298	4.793	2018/05/02
67	175.372411052642	-37.8367003108445	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	67	6.08	2.2	1	14	56.48	On	20180502	0	3	20.16	36.415	0.126	0.014	7.562	7.145	10491.211	9912.643	4.815	2018/05/02
68	175.372418427083	-37.8366809920257	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	68	6.08	2.25	1	16	56.5	On	20180502	0	3	20.56	33.203	0.126	0.013	6.742	6.338	9353.198	8793.113	4.925	2018/05/02
69	175.372427240439	-37.8366612470513	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	69	6.08	2.32	1	19	56.5	On	20180502	0	3	20.658	26.734	0.126	0.01	5.265	4.943	7303.541	6857.722	5.078	2018/05/02
70	175.37243749271	-37.8366420702725	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	70	6.08	2.33	1	21	56.58	On	20180502	0	3	20.67	21.72	0.126	0.008	4.259	3.998	5908.403	5546.907	5.1	2018/05/02
71	175.372449183896	-37.8366234616899	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	71	6.08	2.31	1	25	56.68	On	20180502	0	3	20.67	17.592	0.126	0.007	3.479	3.267	4826.954	4531.624	5.056	2018/05/02
72	175.372462134134	-37.8366059895049	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	72	6.08	2.25	1	29	56.7	On	20180502	0	3	20.394	17.097	0.126	0.007	3.472	3.271	4816.136	4537.199	4.925	2018/05/02
73	175.372475983693	-37.8365907901211	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	73	6.08	2.09	1	34	56.78	On	20180502	0	3	19.46	17.532	0.126	0.007	3.833	3.653	5316.846	5067.678	4.575	2018/05/02
74	175.372490912439	-37.8365768691875	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	74	6.08	2.03	1	38	56.8	On	20180502	0	3	20.29	24.257	0.126	0.009	5.459	5.15	7573.608	7144.288	4.443	2018/05/02
75	175.372506740506	-37.8365638005535	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	75	6.08	2.02	1	42	56.88	On	20180502	0	3	20.592	31.672	0.126	0.012	7.163	6.732	9937.762	9338.909	4.421	2018/05/02
76	175.372519690744	-37.8365541411269	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	76	6.08	1.34	1	45	56.9	On	20180502	0	3	20.664	33.954	0.126	0.013	11.576	10.869	16059.992	15078.527	2.933	2018/05/02
77	175.372502783489	-37.8365623800496	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	77	0.15	0.2	1	44	57.5	On	20180502	0	4	20.552	0.165	0.106	0	15.288	14.374	21209.603	19941.545	0.011	2018/05/02
78	175.372499366066	-37.8365649369565	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	78	0.15	0.34	1	46	57.5	On	20180502	0	4	20.552	0.165	0.106	0	8.993	8.456	12476.237	11730.32	0.018	2018/05/02
79	175.372527784642	-37.8365477488585	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	79	2.43	1.2	1	48	57.32	On	20180502	0	5	20.67	12.714	3.953	0.005	12.111	11.37	16801.846	15773.851	1.05	2018/05/02
80	175.372537677184	-37.8365413565897	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	80	6.08	1.12	1	51	57.3	On	20180502	0	5	19.42	20.624	3.917	0.008	8.413	8.023	11671.488	11130.042	2.451	2018/05/02
81	175.372549188506	-37.8365343961185	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	81	6.08	1.28	1	52	57.3	On	20180502	0	5	20.084	21.825	4.104	0.008	7.79	7.367	10807.169	10220.896	2.802	2018/05/02
82	175.372563397794	-37.8365265833439	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	82	6.08	1.53	1	54	57.3	On	20180502	0	5	20.54	21.57	4.091	0.008	6.441	6.057	8935.582	8402.62	3.349	2018/05/02
83	175.372579045997	-37.8365180603162	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	83	6.08	1.68	1	55	57.3	On	20180502	0	5	20.65	18.568	3.46	0.007	5.05	4.742	7005.158	6578.217	3.677	2018/05/02
84	175.372595773387	-37.8365098213885	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	84	6.08	1.72	1	57	57.3	On	20180502	0	5	20.67	16.331	1.146	0.006	4.338	4.073	6018.08	5649.873	3.765	2018/05/02
85	175.372612680641	-37.8365012983588	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	85	6.08	1.77	1	58	57.3	On	20180502	0	5	20.67	15.836	1.883	0.006	4.088	3.837	5670.701	5323.748	3.874	2018/05/02
86	175.372629587896	-37.8364922071261	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	86	6.08	1.79	1	58	57.3	On	20180502	0	5	19.292	15.476	2.562	0.006	3.95	3.773	5479.78	5233.871	3.918	2018/05/02
87	175.372645056234	-37.8364828317912	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	87	6.08	1.71	1	57	57.3	On	20180502	0	5	18.41	14.11	5.636	0.005	3.77	3.64	5229.85	5049.746	3.743	2018/05/02
88	175.372658186336	-37.8364733144045	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	88	6.08	1.57	1	54	57.3	On	20180502	0	5	18.198	12.549	7.718	0.005	3.652	3.535	5065.988	4904.236	3.436	2018/05/02
89	175.372668438607	-37.8364646493201	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	89	6.08	1.32	1	50	57.3	On	20180502	0	5	18.152	11.768	8.588	0.004	4.073	3.945	5650.666	5473.322	2.889	2018/05/02
90	175.372675992912	-37.8364575467911	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	90	6.08	0.79	1	47	57.22	On	20180502	0	5	18.15	8.751	8.417	0.003	5.061	4.902	7021	6800.815	1.729	2018/05/02
91	175.37267833115	-37.8364551319311	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	91	6.08	0.19	1	44	57.2	On	20180502	0	5	18.15	8.061	6.215	0.003	19.382	18.775	26889.22	26045.949	0.416	2018/05/02
92	175.37263588315	-37.8364984573487	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	92	0.61	1.23	1	56	57	On	20180502	0	5	17.46	4.233	4.406	0.002	15.671	15.308	21740.787	21236.504	0.27	2018/05/02
93	175.372652790404	-37.8364971788941	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	93	1.37	1.91	1	67	57	On	20180502	0	5	17.46	3.062	4.796	0.001	3.251	3.175	4509.593	4404.992	0.942	2018/05/02
94	175.372674913726	-37.8365011563083	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	94	1.98	2.03	1	87	57	On	20180502	0	5	17.46	3.452	5.887	0.001	2.386	2.331	3309.998	3233.222	1.447	2018/05/02
95	175.37269469881	-37.8365130885496	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	95	0.76	2.18	1	113	57	On	20180502	0	5	17.46	4.158	6.418	0.002	6.971	6.809	9670.987	9446.665	0.596	2018/05/02
96	175.37263876098	-37.8365331176645	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	96	0.3	1.32	1	269	57.1	On	20180502	0	6	17.34	0.135	5.742	0	0.948	0.927	1314.645	1286.018	0.143	2018/05/02
97	175.372624731557	-37.8365281458989	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	97	1.06	1.34	1	280	57.1	On	20180502	0	6	13.872	0.48	5.787	0	0.939	0.939	1303.168	1303.168	0.511	2018/05/02
98	175.372572930608	-37.8365129464991	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	98	2.43	1.78	1	288	57.14	On	20180502	0	6	17.34	16.061	5.681	0.006	10.315	10.09	14309.307	13997.72	1.557	2018/05/02
99	175.372552965659	-37.8365084008836	< Unassigned >	harvest_1_Corn_2018_05_02_000 (8220227)	0	99	4.56	1.82	1	287	57.14	On	20180502	0	6	17.34	24.122	6.009	0.009	8.074	7.898	11200.586	10956.692	2.988	2018/05/02

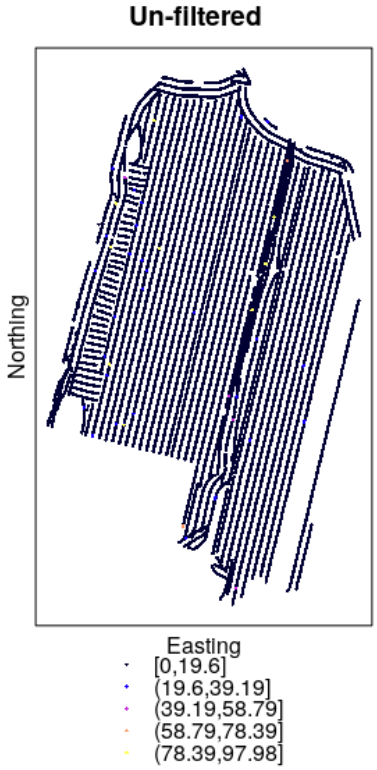
Appendix 2 Customised spatial filtering programme in R

```
##############################################################################
####### Simple program for filtering spatial data points (batch process) #######
##############################################################################

#### load packages

library(readr)
library(sp)
library(rgdal)
library(rgeos)
library(foreach)
library(doParallel)

######## set directory ############

wd <- "E:\\user_name\\input_folder" #--- input yield data file folder
setwd(wd)

boundary_dir = paste0(wd, "/boundary") #---boundary file folder

output_dir = paste0(wd, "/output_data") #--- output data folder

##### read spatial data #####

# read boundary file
bound = readOGR(boundary_dir, as.character(gsub(".shp","",list.files(boundary_dir, pattern = "\\.shp$"))))
proj4string(bound) <- CRS("+init=epsg:4326")
bound_utm <- spTransform(bound, CRS("+init=epsg:2193"))

# read yield file
yield_list <- lapply(list.files(wd, pattern = "\\.shp$"), function(shp_list) {
  layer_name <- as.character(gsub(".shp","",shp_list))
  shp_spdf <-readOGR(dsn = wd, stringsAsFactors = FALSE, verbose = TRUE,
                     useC = TRUE, dropNULLGeometries = TRUE, addCommentsToPolygons = TRUE,
                     layer = layer_name, require_geomType = NULL,
                     p4s = NULL, encoding = 'ESRI Shapefile')
})

for (i in seq(yield_list)) {
  proj4string(yield_list[[i]]) <- CRS("+init=epsg:4326")
  yield_list[[i]] <- spTransform(yield_list[[i]], CRS("+init=epsg:2193"))
  yield_list[[i]] <- yield_list[[i]][bound_utm,]
}

##### filter spatial points #####

#--Global Filter Algorithm--Sudduth et al. 2003
GlobalFilter <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm) # whisker length 1.5 by default
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}
```

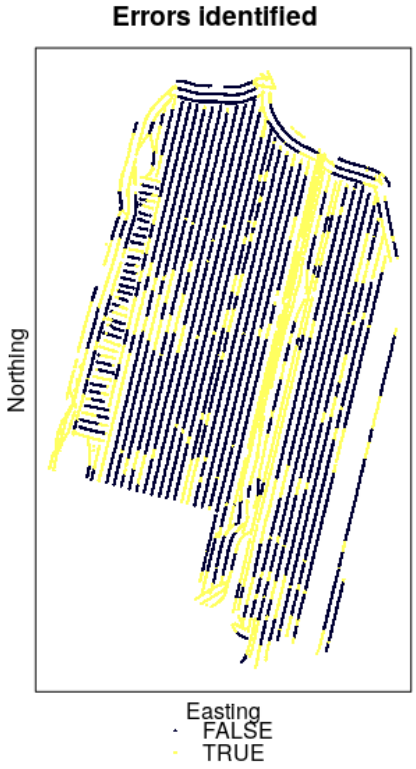
```
YieldData1st <- yield_list

for (i in seq(YieldData1st)) {

 YieldData1st[[i]]$yield <- GlobalFilter(YieldData1st[[i]]$`Yld_Mass_D`)

 # Select yield variable `Yld_Mass_D`, otherwise rename the variable as `Yld_Mass_D`!

 YieldData1st[[i]] <- subset(YieldData1st[[i]], !is.na(YieldData1st[[i]]$yield))

}

#--Local Filter Algorithm--Spekken 2013

YieldData2nd <- YieldData1st

CV <- function(mean, sd) {(sd / mean) * 100}
distThreshold <- 5 # Distance threshold
CVThreshold <- 20 # CV threshold

LocalCV <- list()
Num.CV <- list()

# Parallel processing to reduce processing time
cores=detectCores() #setup parallel backend to use many processors
clust_cores <- makeCluster(cores[1]-1)
registerDoParallel(clust_cores) #To see if the connections are active, use showConnections()

YieldData3rd = foreach(i = seq(YieldData2nd), .combine=list, .multicombine=TRUE) %dopar% {
 LocalCV[[i]] = sapply(X = 1:length(YieldData2nd[[i]]),
                FUN = function(pt) {
                 d = spDistsN1(YieldData2nd[[i]], YieldData2nd[[i]][pt,])
                 ret = CV(mean = mean(YieldData2nd[[i]][d < distThreshold, ]$yield),
                       sd = sd(YieldData2nd[[i]][d < distThreshold, ]$yield))
                 return(ret)
                }) # calculate CV in the local neighbour

 YieldData2nd[[i]]$CV <- LocalCV[[i]]
 YieldData2nd[[i]] <- subset(YieldData2nd[[i]], !is.na(YieldData2nd[[i]]$CV))

 Num.CV[[i]] = sapply(X = 1:length(YieldData2nd[[i]]),
               FUN = function(pt) {
                 d = spDistsN1(YieldData2nd[[i]], YieldData2nd[[i]][pt,])
                 ret = length(YieldData2nd[[i]][d<distThreshold & YieldData2nd[[i]]$CV>CVThreshold,]$CV) ==
length(YieldData2nd[[i]][d<distThreshold,]$CV)
                 return(ret)
                 # If the total number of CVs over 25% equals to the total number of CVs within a search radius then
return TRUE or 1
                }
 )

 YieldData2nd[[i]]$NumCV <- Num.CV[[i]]  # Add num CV as attribute data
 YieldData2nd[[i]] <- subset(YieldData2nd[[i]], YieldData2nd[[i]]$NumCV == FALSE) # subset the filtered data
}

stopCluster(clust_cores)
```

```
##### Save & extract processed data ########

for (i in seq(YieldData3rd)) {

  YieldData3rd[[i]] <- spTransform(YieldData3rd[[i]], CRS("+init=epsg:4326")) # reproject into wgs84

  # writeOGR(YieldData3rd[[i]], output_dir, as.character(gsub(".shp","", list.files(wd, pattern = "\\.shp$")[i])),
driver="ESRI Shapefile")

    # save the processed files to the folder using writeOGR function

    for (i in seq(YieldData3rd))

      assign(as.character(gsub(".shp","", list.files(wd, pattern = "\\.shp$")[i])), YieldData3rd[[i]])
  # Extract each element in list into its own object
  }
```

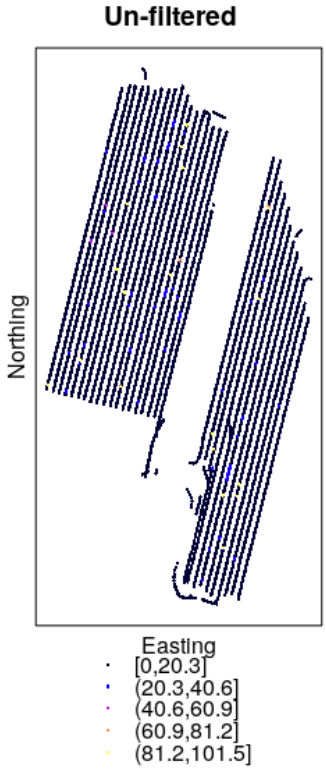
Appendix 3 Locations of filtered data points



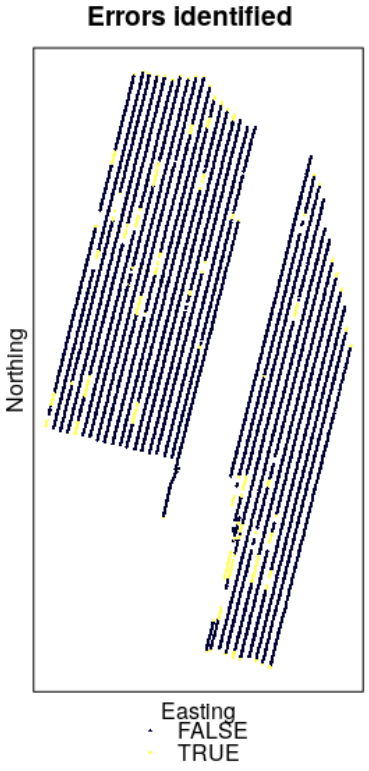
(**a**) 2014 un-filtered yield data



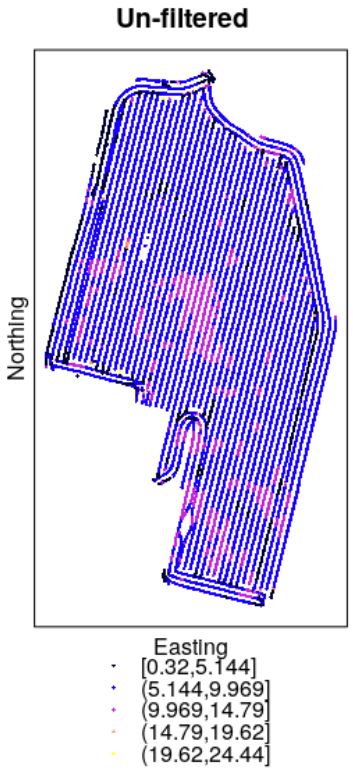
(**b**) 2014 inliers identified



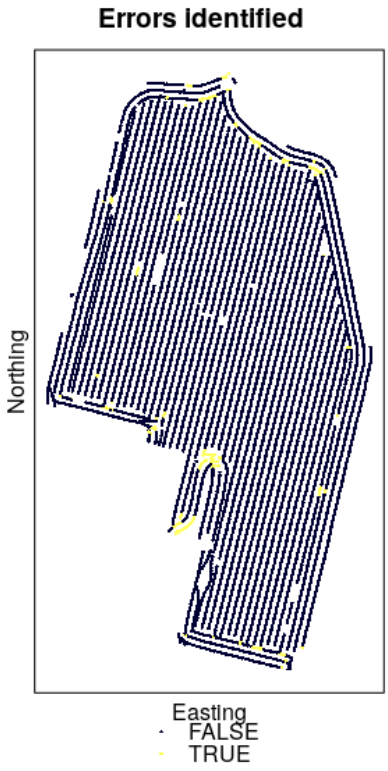
(c) 2015 un-filtered yield data



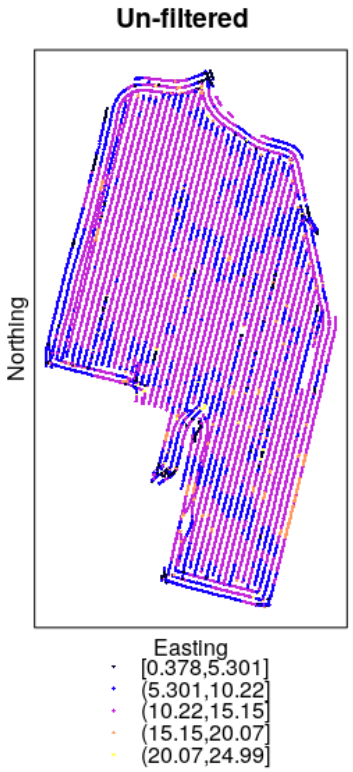
(d) 2015 inliers identified



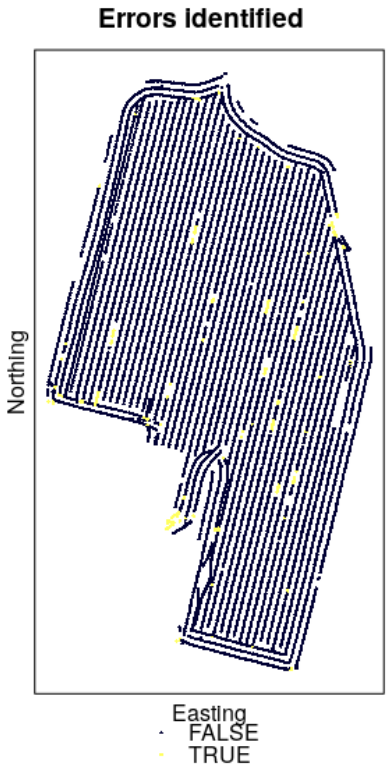
(e) 2017 un-filtered yield data



(f) 2017 inliers identified



(g) 2018 un-filtered yield data



(h) 2018 inliers identified

Figure A - 1 Unfiltered yield monitor data (a, c, e, g) and the locations of the yield measurement errors (b, d, f, h) identified by the filtering programme

Appendix 4 Soil test results for fertiliser recommendation (2018)

Hill Laboratories
TRIED, TESTED AND TRUSTED
R J Hill Laboratories Limited
28 Duke Street Frankton 3204
Private Bag 3205
Hamilton 3240 New Zealand
T 0508 HILL LAB (44 555 22)
T +64 7 858 2000
E mail@hill-labs.co.nz
W www.hill-laboratories.com

Certificate of Analysis

Client:	Foundation for Arable Research	**Lab No:**	2045142	svgpv1
Address:	113C Ruakura Road Hamilton East Hamilton 3214	**Date Received:** **Date Reported:** **Quote No:** **Order No:** **Client Reference:**	07-Sep-2018 12-Sep-2018 1488	
Phone:	03 345 5783	**Submitted By:**	Steve Payne	

Soil Analysis Results

Sample Name:		4	MPT	2a	2b	2c	2d
Lab Number:		2045142.1	2045142.2	2045142.3	2045142.4	2045142.5	2045142.6
Sample Type:		SOIL Maize (150mm)	SOIL Maize (150mm)	SOIL Maize (150mm)	SOIL Maize (150mm)	SOIL Maize (150mm)	SOIL Maize (150mm)
Sample Type Code:		S6	S6	S6	S6	S6	S6
pH	pH Units	6.6	6.4	6.8	7.0	6.9	6.9
Olsen Phosphorus	mg/L	108	47	58	62	85	67
Potassium	me/100g	0.53	0.40	0.42	0.66	0.54	0.75
Potassium	%BS	2.6	1.8	2.2	3.6	2.6	3.8
Potassium	MAF units	9	7	7	11	9	13
Calcium	me/100g	13.4	12.0	13.3	14.2	16.4	14.6
Calcium	%BS	66	55	71	76	80	73
Calcium	MAF units	14	12	14	15	16	15
Magnesium	me/100g	0.69	0.73	0.92	1.13	0.96	0.99
Magnesium	%BS	3.4	3.3	4.9	6.1	4.7	5.0
Magnesium	MAF units	13	13	18	21	17	18
Sodium	me/100g	0.07	0.09	0.06	0.07	0.07	0.08
Sodium	%BS	0.4	0.4	0.3	0.4	0.3	0.4
Sodium	MAF units	3	3	2	3	3	3
CEC	me/100g	20	22	19	19	21	20
Total Base Saturation	%	72	60	79	86	87	82
Volume Weight	g/mL	0.81	0.81	0.86	0.83	0.80	0.82
Potentially Available Nitrogen (15cm Depth)*	kg/ha	28	50	26	38	37	24
Anaerobically Mineralisable N*	µg/g	23	41	20	30	31	19
Soil Sample Depth*	mm	0-150	0-150	0-150	0-150	0-150	0-150

Appendix 5 Soil test results for fertiliser recommendation (2019)

Hill Laboratories
TRIED, TESTED AND TRUSTED
R J Hill Laboratories Limited
28 Duke Street Frankton 3204
Private Bag 3205
Hamilton 3240 New Zealand
T 0508 HILL LAB (44 555 22)
T +64 7 858 2000
E mail@hill-labs.co.nz
W www.hill-laboratories.com

Certificate of Analysis

Client:	Foundation for Arable Research	**Lab No:**	2149187	svgpv1
Address:	113C Ruakura Road Hamilton East Hamilton 3214	**Date Received:** **Date Reported:** **Quote No:** **Order No:** **Client Reference:**	26-Mar-2019 29-Mar-2019 1675 	
Phone:	03 345 5783	**Submitted By:**	Steve Payne	

Soil Analysis Results

Sample Name:		2C	2D	2B	6		
Lab Number:		2149187.1	2149187.2	2149187.3	2149187.4		
Sample Type:		SOIL Arable	SOIL Arable	SOIL Arable	SOIL Arable		
Sample Type Code:		S56	S56	S56	S56		
pH	pH Units	6.5	6.7	7.0	6.3	-	-
Olsen Phosphorus	mg/L	90	66	46	63	-	-
Potassium	me/100g	0.42	1.05	0.59	0.56	-	-
Potassium	%BS	2.2	5.0	2.9	3.5	-	-
Potassium	MAF units	8	20	10	12	-	-
Calcium	me/100g	13.3	14.9	16.2	9.5	-	-
Calcium	%BS	69	71	80	59	-	-
Calcium	MAF units	16	17	17	12	-	-
Magnesium	me/100g	0.84	0.97	1.16	0.81	-	-
Magnesium	%BS	4.3	4.6	5.7	5.0	-	-
Magnesium	MAF units	18	20	21	19	-	-
Sodium	me/100g	0.07	0.08	0.10	0.10	-	-
Sodium	%BS	0.4	0.4	0.5	0.6	-	-
Sodium	MAF units	3	3	4	5	-	-
CEC	me/100g	19	21	20	16	-	-
Total Base Saturation	%	76	81	89	68	-	-
Volume Weight	g/mL	0.94	0.91	0.82	1.02	-	-
Sulphate Sulphur	mg/kg	34	27	14	12	-	-
Potentially Available Nitrogen (15cm Depth)*	kg/ha	83	103	90	101	-	-
Anaerobically Mineralisable N*	µg/g	59	76	73	66	-	-
Soil Sample Depth*	mm	0-150	0-150	0-150	0-150	-	-

Appendix 6 R script for delineating zones

```
#### load packages
library(readr)
library(sp)
library(raster)
library(knitr)
library(rgdal)
library(rgeos)
library(foreach)
library(doParallel)
library(tidyverse)
library(automap)


#### set functions
# batch load shape file
#### set directory
wd <- "~/Documents/test.data/ncrs_back"
setwd(wd)

boundary_dir = paste0(wd, "/boundary")

soil_dir = paste0(wd, "/soilec")

elevation_dir = paste0(wd, "/planting")

rs_dir = paste0(wd, "/rs_data/mz_")


#### load data
# boundary
bound = readOGR(boundary_dir, as.character(gsub(".shp","",list.files(boundary_dir, pattern = "\\.shp$"))))
proj4string(bound) <- CRS("+init=epsg:4326")
bound_utm <- spTransform(bound, CRS("+init=epsg:2193"))

# soilEC
soilEC = readOGR(soil_dir, as.character(gsub(".shp","",list.files(soil_dir, pattern = "\\.shp$"))))
proj4string(soilEC) <- CRS("+init=epsg:4326")
soilEC_utm <- spTransform(soilEC, CRS("+init=epsg:2193"))
soilEC_utm <- soilEC_utm[bound_utm,]

# elevation
elevation = readOGR(elevation_dir, as.character(gsub(".shp","",list.files(elevation_dir, pattern = "\\.shp$")[2])))
proj4string(elevation) <- CRS("+init=epsg:4326")
elevation_utm <- spTransform(elevation, CRS("+init=epsg:2193"))
elevation_utm <- elevation_utm[bound_utm,]

# remote sensing
rs <- lapply(list.files(rs_dir, pattern = ".tif$", full.names = TRUE), stack)
rs_utm <- lapply(rs, function(a){projectRaster(a, crs = CRS("+init=epsg:2193"))})


masked = list()
band_names = c('B','G','R','NIR')
for (i in seq(rs_utm)) {
```

```
  masked[[i]] <- mask(x = rs_utm[[i]], mask = bound_utm)

  for(j in 1:4){
    names(masked[[i]])[j] = paste0(names(masked[[i]])[j], band_names[j])
  }
}


##### yield #####
yield_list <- lapply(list.files(wd, pattern = "\\.shp$"), function(shp_list) {
  layer_name <- as.character(gsub(".shp","",shp_list))
  shp_spdf <-readOGR(dsn = wd, stringsAsFactors = FALSE, verbose = TRUE,
                     useC = TRUE, dropNULLGeometries = TRUE, addCommentsToPolygons = TRUE,
                     layer = layer_name, require_geomType = NULL,
                     p4s = NULL, encoding = 'ESRI Shapefile')
})

for (i in seq(yield_list)) {
  proj4string(yield_list[[i]]) <- CRS("+init=epsg:4326")
  yield_list[[i]] <- spTransform(yield_list[[i]], CRS("+init=epsg:2193"))
  yield_list[[i]] <- yield_list[[i]][bound_utm,]
}

#--Global Filter Algorithm--Sudduth et al. 2003
GlobalFilter <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm) # whisker length 1.5 by default
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}

YieldData1st <- yield_list

for (i in seq(YieldData1st)) {
  YieldData1st[[i]]$yield <- GlobalFilter(YieldData1st[[i]]$`Yld_Mass_D`)
  # Select yield variable `Yld_Mass_D`, otherwise rename the variable as `Yld_Mass_D`!
  YieldData1st[[i]] <- subset(YieldData1st[[i]], !is.na(YieldData1st[[i]]$yield))
} # Only iterate once

#--Local Filter Algorithm--Spekken 2013 (p.s. parallel processing will be used to reduce time usage)

YieldData2nd <- YieldData1st

CV <- function(mean, sd) {(sd / mean) * 100}
distThreshold <- 5 # Distance threshold
CVThreshold <- 20 # CV threshold

LocalCV <- list()
Num.CV <- list()

# Parallel processing
cores=detectCores() #setup parallel backend to use many processors
clust_cores <- makeCluster(cores[1]-1) #not to overload your computer
registerDoParallel(clust_cores) #To see if the connections are active, use showConnections()
```

```
YieldData3rd = foreach(i = seq(YieldData2nd), .combine=list, .multicombine=TRUE) %dopar% {
 LocalCV[[i]] = sapply(X = 1:length(YieldData2nd[[i]]),
                FUN = function(pt) {
                 d = spDistsN1(YieldData2nd[[i]], YieldData2nd[[i]][pt,])
                 ret = CV(mean = mean(YieldData2nd[[i]][d < distThreshold, ]$yield),
                       sd = sd(YieldData2nd[[i]][d < distThreshold, ]$yield))
                 return(ret)
                }) # calculate CV in the local neighbour

 YieldData2nd[[i]]$CV <- LocalCV[[i]]
 YieldData2nd[[i]] <- subset(YieldData2nd[[i]], !is.na(YieldData2nd[[i]]$CV))

 Num.CV[[i]] = sapply(X = 1:length(YieldData2nd[[i]]),
               FUN = function(pt) {
                d = spDistsN1(YieldData2nd[[i]], YieldData2nd[[i]][pt,])
                ret = length(YieldData2nd[[i]][d<distThreshold & YieldData2nd[[i]]$CV>CVThreshold,]$CV) ==
length(YieldData2nd[[i]][d<distThreshold,]$CV)
                # If the total number of CVs over 25% equals to the total number of CVs within a search radius
                return(ret)
                # then return TRUE or 1
               }
 )

 YieldData2nd[[i]]$NumCV <- Num.CV[[i]]
 # Add num CV as attribute data
 YieldData2nd[[i]] <- subset(YieldData2nd[[i]], YieldData2nd[[i]]$NumCV == FALSE)
 # subset the filtered data
}

stopCluster(clust_cores)

################################
# Interpolation using Kriging #
################################

# create a regular grid of 6 m
grd = spsample(bound_utm, type = "regular", cellsize = 6)

# yield mapping
YieldPts <- YieldData3rd

krige_result <- list()
YieldMap <- list()

for (i in seq(YieldPts)) {
 krige_result[[i]] <- autoKrige(yield~1, YieldPts[[i]], grd, maxdist= 20)
 # interpolate yield points into maps using point kriging with maximum search distance of 20m
 YieldMap[[i]] <- krige_result[[i]]$krige_output
 YieldMap[[i]]$yield_norm <- YieldMap[[i]]$var1.pred/mean(YieldMap[[i]]$var1.pred, na.rm =T)
 # normalise yield: dividing each pixel value by the mean
 gridded(YieldMap[[i]]) = T
}


# ---- historical yield trend ----

JustYield <- list()
```

```
for (i in seq(YieldMap)) {
  JustYield[[i]] <- YieldMap[[i]]$yield_norm
}

JustYield.df <- data.frame(t(matrix(unlist(JustYield), nrow=length(JustYield), byrow=T)))
# convert list to data frame

JustCoords <- data.frame(YieldMap[[1]]@coords)

Yield_Average <- rowMeans(JustYield.df,na.rm = T) # average yield
Yield_Std <- apply(JustYield.df,1, sd, na.rm = T) # standard deviation of yield
Yield_CV <- Yield_Std/Yield_Average*100 # coefficient of variation of yield

YieldTrend <- cbind(Yield_Average, Yield_CV, JustCoords)
# put them back as data frame

coordinates(YieldTrend) <- c('x1', 'x2')
# convert to spatial point data frame

YieldTrend$spatial_variability[YieldTrend$Yield_Average > 1] <- 'HY'
YieldTrend$spatial_variability[YieldTrend$Yield_Average < 1] <- 'LY'


gridded(YieldTrend) <- T
yield_zones <- raster(YieldTrend["spatial_variability"])
# writeRaster(yield_zones,'arithmetic_yield.tif',options=c('TFW=YES'), overwrite=TRUE)

##### soil/elevation attribute mapping #####

elevation_map <- autoKrige(Elvtn__~1, elevation_utm, grd, maxdist= 20)
elevationMap <- elevation_map$krige_output


soilom_map <- autoKrige(Soil_OM___~1, soilEC_utm, grd, maxdist= 20)
soilomMap <- soilom_map$krige_output

soilec_shallow_map <- autoKrige(EC_Shallow ~1, soilEC_utm, grd, maxdist= 20)
soilec_shallowMap <- soilec_shallow_map$krige_output

soilec_deep_map <- autoKrige(EC_Deep_dS~1, soilEC_utm, grd, maxdist= 20)
soilec_deepMap <- soilec_deep_map$krige_output

soil_df = data.frame(soilECdeep = soilec_deepMap$var1.pred, soilECshallow = soilec_shallowMap$var1.pred)

soil_PCs <- prcomp(soil_df, center = T, scale = T) # PCA

soil_clusters_output = cmeans(soil_PCs$x[,1:2], 2, iter.max = 10000, dist = "euclidean", method = "cmeans", m = 2)

grd$soil_clusters = soil_clusters_output$cluster
gridded(grd) <- T
soil_zones <- raster(grd["soil_clusters"])
# writeRaster(soil_zones,'soil_zone.tif', options=c('TFW=YES'), overwrite=TRUE)

plot(soil_zones, col = c("grey70", "grey30"), legend = FALSE, main = "Soil zones")

##### multispectral reflectance analysis #####
```

```
.rs.unloadPackage("tidyr") # package conflict to tidyr, which also has extract function

ref = list()
for (i in seq(masked)) {
  ref[[i]] = data.frame(extract(masked[[i]], grd))
}

reflect = na.omit(cbind(grd@coords, do.call(data.frame, ref)))

reflect_PCs = prcomp(reflect[,3:ncol(reflect)], center = T, scale = T) # PCA
reflect_clusters_output = cmeans(reflect_PCs$x[,1:2], 2, iter.max = 10000, dist = "euclidean", method = "cmeans",
m = 2)

reflect$reflect_clusters = reflect_clusters_output$cluster
coordinates(reflect) = c('x1','x2')

gridded(reflect) <- T
reflect_zones <- raster(reflect["reflect_clusters"])
# writeRaster(reflect_zones,'reflect_zone.tif', options=c('TFW=YES'), overwrite=TRUE)

##### Multiple-year yield analysis #####

soil = list()
yd = list()
refl = list()
for (i in seq(YieldData3rd)) {

  soil[[i]] = data.frame(extract(soil_zones, YieldData3rd[[i]]))
  yd[[i]] = data.frame(extract(yield_zones, YieldData3rd[[i]]))
  refl[[i]] = data.frame(extract(reflect_zones, YieldData3rd[[i]]))

  }


NCRS = cbind(NCRS, soilMZs_code = unlist(soil), yieldMZs_code = unlist(yd), reflectMZs_code = unlist(refl),
          soilMZs = NA, yieldMZs = NA, reflectMZs = NA)

if(mean(NCRS[which(NCRS$soilMZs_code==1),]$yield, na.rm =T) >
mean(NCRS[which(NCRS$soilMZs_code==2),]$yield, na.rm =T)){
  NCRS[which(NCRS$soilMZs_code==1),]$soilMZs = 'HY'
  NCRS[which(NCRS$soilMZs_code==2),]$soilMZs = 'LY'
} else {
  NCRS[which(NCRS$soilMZs_code==1),]$soilMZs = 'LY'
  NCRS[which(NCRS$soilMZs_code==2),]$soilMZs = 'HY'
}

if(mean(NCRS[which(NCRS$yieldMZs_code==1),]$yield, na.rm =T) >
mean(NCRS[which(NCRS$yieldMZs_code==2),]$yield, na.rm =T)){
  NCRS[which(NCRS$yieldMZs_code==1),]$yieldMZs = 'HY'
  NCRS[which(NCRS$yieldMZs_code==2),]$yieldMZs = 'LY'
} else {
  NCRS[which(NCRS$yieldMZs_code==1),]$yieldMZs = 'LY'
  NCRS[which(NCRS$yieldMZs_code==2),]$yieldMZs = 'HY'
}

if(mean(NCRS[which(NCRS$reflectMZs_code==1),]$yield, na.rm =T) >
mean(NCRS[which(NCRS$reflectMZs_code==2),]$yield, na.rm =T)){
```

```
  NCRS[which(NCRS$reflectMZs_code==1),]$reflectMZs = 'HY'
  NCRS[which(NCRS$reflectMZs_code==2),]$reflectMZs = 'LY'
} else {
  NCRS[which(NCRS$reflectMZs_code==1),]$reflectMZs = 'LY'
  NCRS[which(NCRS$reflectMZs_code==2),]$reflectMZs = 'HY'
}

NCRS$soilMZs = factor(NCRS$soilMZs)
NCRS$yieldMZs = factor(NCRS$yieldMZs)
NCRS$reflectMZs = factor(NCRS$reflectMZs)

grd$yield_clusters = extract(yield_zones, grd)
grd$reflect_clusters = extract(reflect_zones, grd)
```

Appendix 7 R script for combining data

```
#### load packages
library(readr)
library(sp)
library(raster)
library(knitr)
library(rgdal)
library(rgeos)
library(foreach)
library(doParallel)
library(tidyverse)
library(automap)

#### set functions
# batch load shape file

# extract spatial pt info within nearest neighbor
extract_sppt_id <- function(set1, set2){
  set1sp <- SpatialPoints(set1)
  set2sp <- SpatialPoints(set2)
  apply(gDistance(set2sp, set1sp, byid=T), 1, which.min)
}

#### set directory
wd <- "~/Documents/test.data/ncrs_back"
setwd(wd)

boundary_dir = paste0(wd, "/boundary/boundary_hdland_excl")

soil_dir = paste0(wd, "/soilec")

elevation_dir = paste0(wd, "/planting")

weather_dir = paste0(wd, "/weather")


#### load data
# boundary
bound = readOGR(boundary_dir, as.character(gsub(".shp","",list.files(boundary_dir, pattern = "\\.shp$"))))
proj4string(bound) <- CRS("+init=epsg:4326")
bound_utm <- spTransform(bound, CRS("+init=epsg:2193"))

# soil
soilEC = readOGR(soil_dir, as.character(gsub(".shp","",list.files(soil_dir, pattern = "\\.shp$"))))
proj4string(soilEC) <- CRS("+init=epsg:4326")
soilEC_utm <- spTransform(soilEC, CRS("+init=epsg:2193"))

soilEC_utm <- soilEC_utm[bound_utm,]

# elevation
elevation = readOGR(elevation_dir, as.character(gsub(".shp","",list.files(elevation_dir, pattern = "\\.shp$")[2])))
proj4string(elevation) <- CRS("+init=epsg:4326")
elevation_utm <- spTransform(elevation, CRS("+init=epsg:2193"))
elevation_utm <- elevation_utm[bound_utm,]
```

```
# weather
weather = read_csv(paste0(weather_dir, "/weather20130901onward"))

 weather$`Tmax(C)` = as.numeric(weather$`Tmax(C)`)
 weather$`Tmin(C)` = as.numeric(weather$`Tmin(C)`)
 weather$GDD = (weather$`Tmin(C)` + weather$`Tmax(C)`)/2 -8
 weather$GDD[weather$GDD < 0] = 0

weather_df = data.frame(weather[c("Day(Local_Date)","Rain(mm)","GDD", "Rad(MJ/m2)")] )


weather_df$Rain.mm. = as.numeric(weather_df$Rain.mm.)
weather_df$Rad.MJ.m2. = as.numeric(weather_df$Rad.MJ.m2.)
weather_df$Day.Local_Date. = as.Date(weather_df$Day.Local_Date., format = "%Y%m%d:0000")

##### remote sensing
 rs <- lapply(list.files(rs_dir, pattern = ".tif$", full.names = TRUE), stack)
 rs_utm <- lapply(rs, function(a){projectRaster(a, crs = CRS("+init=epsg:2193"))})

 savi = list()
 masked = list()
 for (i in seq(rs_utm)) {
  savi[[i]] = 1.5*(rs_utm[[i]][[4]]-rs_utm[[i]][[3]])/(rs_utm[[i]][[4]]+rs_utm[[i]][[3]]+0.5)
  masked[[i]] <- mask(x = savi[[i]], mask = bound_utm)
 }

# yield

yield_list <- lapply(list.files(wd, pattern = "\\.shp$"), function(shp_list) {
  layer_name <- as.character(gsub(".shp","",shp_list))
  shp_spdf <-readOGR(dsn = wd, stringsAsFactors = FALSE, verbose = TRUE,
                 useC = TRUE, dropNULLGeometries = TRUE, addCommentsToPolygons = TRUE,
                 layer = layer_name, require_geomType = NULL,
                 p4s = NULL, encoding = 'ESRI Shapefile')
})

for (i in seq(yield_list)) {
  proj4string(yield_list[[i]]) <- CRS("+init=epsg:4326")
  yield_list[[i]] <- spTransform(yield_list[[i]], CRS("+init=epsg:2193"))
  yield_list[[i]] <- yield_list[[i]][bound_utm,]
}

#--Global Filter Algorithm--Sudduth et al. 2003
GlobalFilter <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm) # whisker length 1.5 by default
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}

YieldData1st <- yield_list

for (i in seq(YieldData1st)) {
  YieldData1st[[i]]$yield <- GlobalFilter(YieldData1st[[i]]$`Yld_Mass_D`)
  # Select yield variable `Yld_Mass_D`, otherwise rename the variable as `Yld_Mass_D`!
```

```
  YieldData1st[[i]] <- subset(YieldData1st[[i]], !is.na(YieldData1st[[i]]$yield))
} # Only iterate once

#--Local Filter Algorithm--Spekken 2013 (p.s. parallel processing will be used to reduce time usage)

YieldData2nd <- YieldData1st

CV <- function(mean, sd) {(sd / mean) * 100}
distThreshold <- 5 # Distance threshold
CVThreshold <- 20 # CV threshold

LocalCV <- list()
Num.CV <- list()

# Parallel processing
cores=detectCores() #setup parallel backend to use many processors
clust_cores <- makeCluster(cores[1]-1) #not to overload your computer
registerDoParallel(clust_cores) #To see if the connections are active, use showConnections()

YieldData3rd = foreach(i = seq(YieldData2nd), .combine=list, .multicombine=TRUE) %dopar% {
  LocalCV[[i]] = sapply(X = 1:length(YieldData2nd[[i]]),
                  FUN = function(pt) {
                    d = spDistsN1(YieldData2nd[[i]], YieldData2nd[[i]][pt,])
                    ret = CV(mean = mean(YieldData2nd[[i]][d < distThreshold, ]$yield),
                          sd = sd(YieldData2nd[[i]][d < distThreshold, ]$yield))
                    return(ret)
                  }) # calculate CV in the local neighbour

  YieldData2nd[[i]]$CV <- LocalCV[[i]]
  YieldData2nd[[i]] <- subset(YieldData2nd[[i]], !is.na(YieldData2nd[[i]]$CV))

  Num.CV[[i]] = sapply(X = 1:length(YieldData2nd[[i]]),
                  FUN = function(pt) {
                    d = spDistsN1(YieldData2nd[[i]], YieldData2nd[[i]][pt,])
                    ret = length(YieldData2nd[[i]][d<distThreshold & YieldData2nd[[i]]$CV>CVThreshold,]$CV) ==
length(YieldData2nd[[i]][d<distThreshold,]$CV)
                    # If the total number of CVs over 25% equals to the total number of CVs within a search radius
                    return(ret)
                    # then return TRUE or 1
                  }
  )

  YieldData2nd[[i]]$NumCV <- Num.CV[[i]]
  # Add num CV as attribute data
  YieldData2nd[[i]] <- subset(YieldData2nd[[i]], YieldData2nd[[i]]$NumCV == FALSE)
  # subset the filtered data
}

stopCluster(clust_cores)


################################
# Interpolation using Kriging #
################################

# create a regular grid of 6 m
grd = spsample(bound_utm, type = "regular", cellsize = 24)
```

```
# yield mapping
YieldPts <- YieldData3rd

krige_result <- list()
YieldMap <- list()

for (i in seq(YieldPts)) {
  krige_result[[i]] <- autoKrige(yield~1, YieldPts[[i]], grd, maxdist= 50)
  # interpolate yield points into maps using point kriging with maximum search distance of 20m
  YieldMap[[i]] <- krige_result[[i]]$krige_output
  YieldMap[[i]]$yield_norm <- YieldMap[[i]]$var1.pred/mean(YieldMap[[i]]$var1.pred, na.rm =T)
  # normalise yield: dividing each pixel value by the mean
  gridded(YieldMap[[i]]) = T
}

##### soil/elevation attribute mapping
elevation_map <- autoKrige(Elvtn__~1, elevation_utm, grd, maxdist= 50)
soilom_map <- autoKrige(Soil_OM___~1, soilEC_utm, grd, maxdist= 50)
soilec_shallow_map <- autoKrige(EC_Shallow ~1, soilEC_utm, grd, maxdist= 50)
soilec_deep_map <- autoKrige(EC_Deep_dS~1, soilEC_utm, grd, maxdist= 50)

elevationMap <- elevation_map$krige_output
soilomMap <- soilom_map$krige_output
soilec_shallowMap <- soilec_shallow_map$krige_output
soilec_deepMap <- soilec_deep_map$krige_output

#### join soil/elevation to yield

df = list()
for (i in seq(YieldData3rd)) {

  YieldMap[[i]]$soilec_shallow = soilec_shallowMap$var1.pred
  YieldMap[[i]]$soilec_deep = soilec_deepMap$var1.pred
  YieldMap[[i]]$elevation = elevationMap$var1.pred
  YieldMap[[i]]$soilom = soilomMap$var1.pred


  df[[i]] =  data.frame(year = gsub(".shp","",list.files(wd, pattern = "\\.shp$"))[i],

                        YieldMap[[i]]@coords,
                        yield = YieldMap[[i]]$var1.pred,
                        soilec_shallow = YieldMap[[i]]$soilec_shallow,
                        soilec_deep = YieldMap[[i]]$soilec_deep,
                        elevation = YieldMap[[i]]$elevation,
                        soilom = YieldMap[[i]]$soilom

                        )

  assign(as.character(gsub(".shp","", list.files(wd, pattern = "\\.shp$")[i])), df[[i]])
  }

#### join weather
# seperate dates for yield datasets

season2014_planting_date = '2013-10-08'
season2014_harvest_date = '2014-05-12'
```

```
season2014 = weather_df[(which(weather_df$Day.Local_Date.== season2014_planting_date)-
7):which(weather_df$Day.Local_Date.== season2014_harvest_date),]
season2014 =
  season2014 %>%
  group_by(week = cut(Day.Local_Date., "7 days")) %>%
  summarise(rain=sum(Rain.mm., na.rm = T), rad=sum(Rad.MJ.m2., na.rm = T), GDD= sum(GDD, na.rm = T))
for (i in seq(NROW(season2014))) {season2014$day[i] = as.integer(difftime(season2014$week[i],
season2014$week[1]))-7}

season2015_planting_date = '2014-10-10'
season2015_harvest_date = '2015-05-18'
season2015 = weather_df[(which(weather_df$Day.Local_Date.== season2015_planting_date)-
7):which(weather_df$Day.Local_Date.== season2015_harvest_date),]
season2015 =
  season2015 %>%
  group_by(week = cut(Day.Local_Date., "7 days")) %>%
  summarise(rain=sum(Rain.mm., na.rm = T), rad=sum(Rad.MJ.m2., na.rm = T), GDD= sum(GDD, na.rm = T))
for (i in seq(NROW(season2015))) {season2015$day[i] = as.integer(difftime(season2015$week[i],
season2015$week[1]))-7}

season2017_planting_date = '2016-10-17'
season2017_harvest_date = '2017-05-23'
season2017 = weather_df[(which(weather_df$Day.Local_Date.== season2017_planting_date)-
7):which(weather_df$Day.Local_Date.== season2017_harvest_date),]
season2017 =
  season2017 %>%
  group_by(week = cut(Day.Local_Date., "7 days")) %>%
  summarise(rain=sum(Rain.mm., na.rm = T), rad=sum(Rad.MJ.m2., na.rm = T), GDD= sum(GDD, na.rm = T))
for (i in seq(NROW(season2017))) {season2017$day[i] = as.integer(difftime(season2017$week[i],
season2017$week[1]))-7}

season2018_planting_date = '2017-10-18'
season2018_harvest_date = '2018-05-14'
season2018 = weather_df[(which(weather_df$Day.Local_Date.== season2018_planting_date)-
7):which(weather_df$Day.Local_Date.== season2018_harvest_date),]
season2018 = season2018 %>%
  group_by(week = cut(Day.Local_Date., "7 days")) %>%
  summarise(rain=sum(Rain.mm., na.rm = T), rad=sum(Rad.MJ.m2., na.rm = T), GDD= sum(GDD, na.rm = T))
for (i in seq(NROW(season2018))) {season2018$day[i] = as.integer(difftime(season2018$week[i],
season2018$week[1]))-7}


# "Rain"
ncrs2014df_join = cbind(ncrsgrain2014, setNames(as.list(season2014$rain), paste0("Rain", season2014$day)))
ncrs2015df_join = cbind(ncrsgrain2015, setNames(as.list(season2015$rain), paste0("Rain", season2015$day)))
ncrs2017df_join = cbind(ncrsgrain2017, setNames(as.list(season2017$rain), paste0("Rain", season2017$day)))
ncrs2018df_join = cbind(ncrsgrain2018, setNames(as.list(season2018$rain), paste0("Rain", season2018$day)))

# "radiation"
ncrs2014df_join = cbind(ncrs2014df_join, setNames(as.list(season2014$rad), paste0("rad", season2014$day)))
ncrs2015df_join = cbind(ncrs2015df_join, setNames(as.list(season2015$rad), paste0("rad", season2015$day)))
ncrs2017df_join = cbind(ncrs2017df_join, setNames(as.list(season2017$rad), paste0("rad", season2017$day)))
ncrs2018df_join = cbind(ncrs2018df_join, setNames(as.list(season2018$rad), paste0("rad", season2018$day)))

# "GDD"
ncrs2014df_join = cbind(ncrs2014df_join, setNames(as.list(season2014$GDD), paste0("GDD", season2014$day)))
ncrs2015df_join = cbind(ncrs2015df_join, setNames(as.list(season2015$GDD), paste0("GDD", season2015$day)))
```

```
ncrs2017df_join = cbind(ncrs2017df_join, setNames(as.list(season2017$GDD), paste0("GDD", season2017$day)))
ncrs2018df_join = cbind(ncrs2018df_join, setNames(as.list(season2018$GDD), paste0("GDD", season2018$day)))

NCRS = bind_rows(ncrs2014df_join, ncrs2015df_join, ncrs2017df_join, ncrs2018df_join)
```

Appendix 8 R script for model evaluation

```
#### load packages
library(readr)
library(sp)
library(raster)
library(knitr)
library(rgdal)
library(rgeos)
library(foreach)
library(doParallel)
library(tidyverse)
library(caret)


#### set directory
wd <- "~/Documents/test.data/join_data"
setwd(wd)

#### Data mining
NCRS_weekly <- read.csv("~/Documents/test.data/join_data/pooled.csv")
NCRS = na.omit(NCRS_weekly[, -which(colnames(NCRS_weekly) %in% c('X'))])
NCRS = na.omit(NCRS[, which(colnames(NCRS) %in%
c("year","x1","x2","yield","soilec_shallow","soilec_deep","elevation","soilom",
                                    "Rain.7","Rain0","Rain7","Rain14","Rain21","Rain28","Rain35",
                                    "rad.7","rad0","rad7","rad14","rad21","rad28","rad35",
                                    "GDD.7","GDD0","GDD7","GDD14","GDD21","GDD28","GDD35"))])

# Parallel processing
cores=detectCores()
clust_cores <- makeCluster(cores[1]-1)
registerDoParallel(clust_cores)

# multicolinearity
 corMatrix = cor(NCRS[,-c(1:4)], method = 'pearson')
 highlyCor = findCorrelation(corMatrix,cutoff = 0.8)
 cor_rm = cbind(NCRS[,-c(1:4)][,-highlyCor], yield=NCRS[,4])

 fit_control <- trainControl(method = "repeatedcv", number = 10, repeats = 1)

set.seed(29)
idx_train <- sample(1:nrow(cor_rm), size = nrow(cor_rm)*0.75)

training <- cor_rm[ idx_train, ]
testing  <- cor_rm[ -idx_train, ]

fit_lm =  train(yield ~., data = training, method  = 'lm', trControl = fit_control)
summary(fit_lm)


pred <- predict(fit_lm, testing)
error = postResample(pred = pred, obs = testing$yield)


##############################
###### neural network #######
```

```
############################

set.seed(29)

start_time <- Sys.time() # measure running time
fit_nn <- train(

  x = training[,-NCOL(training)],
  y = training$yield/25,

  method = "nnet", trControl = fit_control,
  tuneGrid = expand.grid(size = 1:5, decay = c(0.1, 0.2, 0.3)),
  linout = TRUE, maxit = 500)

end_time <- Sys.time() # measure running time
time_nn = end_time - start_time # 2.374481 hours

fit_nn
plot(fit_nn)

pred_nn <- predict(fit_nn, testing)
error_nn = postResample(pred = pred_nn*25, obs = testing$yield)


######################
###### rpart2 #######
######################

cor_rm2 = cbind(NCRS[,-c(1:4)], yield=NCRS[,4])

training2 <- cor_rm2[ idx_train, ]
testing2  <- cor_rm2[ -idx_train, ]

fit_ctr <- train(

  x = training2[,-NCOL(training2)],
  y = training2$yield,

  method = "rpart2",
  trControl = fit_control,
  tuneGrid = expand.grid(maxdepth=1:20)
)

fit_ctr
plot(fit_ctr)

pred_ctr <- predict(fit_ctr, testing2)
error_ctr = postResample(pred = pred_ctr, obs = testing2$yield)

######################
####### Cubist ######
######################

set.seed(29)

start_time <- Sys.time() # measure running time
```

```
fit_cubist <- train(

  x = training2[,-NCOL(training2)],
  y = training2$yield,

  method = "cubist",
  tuneGrid = expand.grid(committees = c(1, 5, 10, 20), neighbors = c(0, 5, 9)),
  trControl = fit_control
)

end_time <- Sys.time() # measure running time
time_cu = end_time - start_time # 3.948066 mins

fit_cubist
plot(fit_cubist)
varImp(fit_cubist)

pred_cubist <- predict(fit_cubist, testing2)
error_cubist = postResample(pred = pred_cubist, obs = testing2$yield)


############################
####### Random forest ######
############################

set.seed(29)

start_time <- Sys.time() # measure running time

fit_rf <- train(

  x = training2[,-NCOL(training2)],
  y = training2$yield,

  method = "rf",
  trControl = fit_control,
  importance = TRUE
)

end_time <- Sys.time() # measure running time
time_rf = end_time - start_time #

fit_rf
plot(fit_rf)
varImp(fit_rf)

pred_rf <- predict(fit_rf, testing2)
error_rf = postResample(pred = pred_rf, obs = testing2$yield)


###### Hold out one site/year analysis (hooya) for Random Forest

hooya_rf = cbind(year = NCRS[,1], cor_rm2)

hooya_training5 = list()
hooya_testing5 = list()
hooya_fit_rf = list()
```

```
hooya_pred_rf = list()
hooya_error_rf = list()

names_sites = c("ncrsgrain2014", "ncrsgrain2015", "ncrsgrain2017", "ncrsgrain2018",
            "bell2014_F2", "bell2015_F2" ,  "bell2016_F2",  "bell2017_F2" ,"bell2018_F2",
            "bell2014" ,  "bell2015",  "bell2016", "bell2017", "bell2018",
            "kaipograin2005", "kaipograin2007", "kaipograin2009", "kaipograin2010", "kaipograin2013",
"kaipograin2015", "kaipograin2017_non_trial" ,
            "stanley2008", "stanley2009", "stanley2010")

start_time <- Sys.time() # measure running time
for (i in seq(nlevels(hooya_rf[,1]))) {

  hooya_testing5[[i]]  = subset(hooya_rf, year == names_sites[i])
  hooya_training5[[i]] = subset(hooya_rf, year != names_sites[i])

  set.seed(29)

  hooya_fit_rf[[i]] <- train(

    x = hooya_training5[[i]][, -c(1,ncol(hooya_training5[[i]]))],
    y = hooya_training5[[i]]$yield,

    method = "rf",
    trControl = fit_control
  )

  hooya_pred_rf[[i]] <- predict(hooya_fit_rf[[i]], hooya_testing5[[i]])
  hooya_error_rf[[i]] = postResample(pred = hooya_pred_rf[[i]], obs = hooya_testing5[[i]]$yield)
}

end_time <- Sys.time() # measure running time
time_rf2 = end_time - start_time # 9.698936 mins



######################
####### xgboost ######
######################

set.seed(29)

start_time <- Sys.time() # measure running time

fit_xg <- train(

  x = training2[,-NCOL(training2)],
  y = training2$yield,

  method = "xgbTree",
  trControl = fit_control
)

end_time <- Sys.time() # measure running time
time_xg = end_time - start_time #

fit_xg
```

```
plot(fit_xg)


pred_xg <- predict(fit_xg, testing2)
error_xg = postResample(pred = pred_xg, obs = testing2$yield)


###### Hold out one year analysis (hooya) for xgboost

hooya_xg = cbind(year = NCRS[,1], cor_rm2)

hooya_training6 = list()
hooya_testing6 = list()
hooya_fit_xg = list()
hooya_pred_xg = list()
hooya_error_xg = list()

start_time <- Sys.time() # measure running time
for (i in seq(nlevels(hooya_xg[,1]))) {

 hooya_testing6[[i]]  = subset(hooya_xg, year == names_sites[i])
 hooya_training6[[i]] = subset(hooya_xg, year != names_sites[i])

 set.seed(29)

 hooya_fit_xg[[i]] <- train(

  x = hooya_training6[[i]][, -c(1,ncol(hooya_training6[[i]]))],
  y = hooya_training6[[i]]$yield,

  method = "xgbTree",
  trControl = fit_control
 )

 hooya_pred_xg[[i]] <- predict(hooya_fit_xg[[i]], hooya_testing6[[i]])
 hooya_error_xg[[i]] = postResample(pred = hooya_pred_xg[[i]], obs = hooya_testing6[[i]]$yield)
}

end_time <- Sys.time() # measure running time
time_xg2 = end_time - start_time # 6.155648 hours


stopCluster(clust_cores) # close connection
```