A DATA DRIVEN APPROACH FOR DIAGNOSIS AND MANAGEMENT OF YIELD VARIABILITY ATTRIBUTED TO SOIL CONSTRAINTS

A Thesis submitted by

Stirling Donoghue Roberton, BEng (Agricultural) Hons

For the award of

Doctor of Philosophy

2019

Dedicated to my loving mother and father, Merilyn and Don (1943 – 1995). Thank you for this wonderful opportunity

Abstract

Australian agriculture does not value data to the level required for true precision management. Consequently, agronomic recommendations are frequently based on limited soil information and do not adequately address the spatial variance of the constraints presented. This leads to lost productivity. Due to the costs of soil analysis, land owners and practitioners are often reluctant to invest in soil sampling exercises as the likely economic gain from this investment has not been adequately investigated. A value proposition is therefore required to realise the agronomic and economic benefits of increased site-specific data collection with the aim of ameliorating soil constraints. This study is principally concerned with identifying this value proposition by investigating the spatially variable nature of soil constraints and their interactions with crop yield at the sub-field scale. Agronomic and economic benefits are quantified against simulated ameliorant recommendations made on the basis of varied sampling approaches.

In order to assess the effects of sampling density on agronomic recommendations, a 108 ha site was investigated, where 1200 direct soil measurements were obtained (300 sample locations at 4 depth increments) to form a benchmark dataset for analysis used in this study. Random transect sampling (for field average estimates), zone management, regression kriging (SSPFe) and ordinary kriging approaches were first investigated at various sampling densities (N=10, 20, 50, 100, 150, 200, 250 and 300) to observe the effects of lime and gypsum ameliorant recommendation advice. It was identified that the ordinary kriging method provided the most accurate spatial recommendation advice for gypsum and lime at all depth increments investigated (i.e. 0–10 cm, 10–20 cm, 20–40 cm and 40–60 cm), with the majority of improved accuracy being achieved up to 50 samples (≈0.5 samples/ha). The lack of correlation between the environmental covariates and target soil variables inhibited the ability for regression kriging to outperform ordinary kriging.

To extend these findings in an attempt to identify the economically optimal sampling density for the investigation site, a yield prediction model was required to estimate the spatial yield response due to amelioration. Given the complex nonlinear relationships between soil properties and yield, this was achieved by applying four machine learning models (both linear and nonlinear) consisting of a mixed-linear regression, a regression tree (Cubist), an artificial neural network and a support vector machine. These were trained using the 1200 directly measured soil samples, each with 9 soil measurements describing structural features (i.e. soil pH, exchangeable sodium

percentage, electrical conductivity, clay, silt, sand, bulk density, potassium, cation exchange capacity) to predict the spatial yield variability at the investigation site with four years of yield data. It was concluded that the Cubist regression tree model produced superior results in terms of improved generalization, whilst achieving an acceptable R^2 for training and validation (up to R^2=0.80 for training and R^2=0.78 for validation). The lack of temporal yield information constrained the ability to develop a temporally stable yield prediction model to account for the uncertainties of climate interactions associated with the spatial variability of yield. Accurate predictive performance was achieved for single-season models.

Of the spatial prediction methods investigated, random transect sampling and ordinary kriging approaches were adopted to simulate 'blanket-rate' (BR) and 'variable-rate' (VR) gypsum applications, respectively, for the amelioration of sodicity at the investigated site. For each sampling density, the spatial yield response as a result of a BR and VR application of gypsum was estimated by application of the developed Cubist yield prediction model, calibrated for the investigation site. Accounting for the cost of sampling and financial gains, due to a yield response, the most economically optimum sampling density for the investigation site was 0.2 cores/ha for 0–20 cm treatment and 0.5 cores/ha for 0–60 cm treatment taking a VR approach. Whilst this resulted in an increased soil data investment of $26.4/ha and $136/ha for 0–20 cm and 0–60 cm treatment respectively in comparison to a BR approach, the yield gains due to an improved spatial gypsum application were in excess of 6 t and 26 t per annum. Consequently, the net benefit of increased data investment was estimated to be up to $104,000 after 20 years for 0–60 cm profile treatment.

Identifying the influence on qualitative data and management information on soil-yield interaction, a probabilistic approach was investigated to offer an alternative approach where empirical models fail. Using soil compaction as an example, a Bayesian Belief Network was developed to explore the interactions of machine loading, soil wetness and site characteristics with the potential yield declines due to compaction induced by agricultural traffic. The developed tool was subsequently able to broadly describe the agronomic impacts of decisions made in data limiting environments.

This body of work presents a combined approach to improving both the diagnosis and management of soil constraints using a data driven approach. Subsequently, a detailed discussion

is provided to further this work, and improve upon the results obtained. By continuing this work it is possible to change the industry attitude to data collection and significantly improve the productivity, profitability and soil husbandry of agricultural systems.

Certification of Thesis

This Thesis is entirely the work of Stirling Roberton except where otherwise acknowledged. The work is original and has not previously been submitted for any other award, except where acknowledged.

Principal Supervisor: Associate Prof. John McLean Bennett

Associate Supervisor: Dr Craig Lobsey

Associate Supervisor: Prof. Yan Li

Student and supervisors signatures of endorsement are held at the University.

Acknowledgements

First and foremost I would like to thank my principal supervisor, Associate Prof. John McLean Bennett for all his efforts over the past few years in helping this thesis come to fruition. Your countless hours of editing and guidance have not gone unnoticed, and I am grateful for this. I thank you for our thought provoking discussions, which have often lead to the discovery of new and exciting ideas. John's motivation to help people achieved their potential is a great quality of his, and is one that is appreciated by all his students, including myself.

Thank you to my associate supervisors, Prof. Yan Li and Dr Craig Lobsey for your guidance and suggestions over the past few years. I look forward to our future discussions in progressing the development of machine learning within agriculture and pedometrics.

I am greatly appreciative for all the other staff at the University of Southern Queensland (USQ) who have assisted me during this journey. To Dr Alla Marchuk and Dr Susette Eberhard, thank you for patiently teaching me sound laboratory protocols and procedures that assisted in the efficient processing of my 1200 samples. A special thank you to Ian Grant from Agricultural Chemistry for your guidance in the analysis of soil chemistry. Ian has an invaluable wealth of experience and I am grateful to have spent some time with him learning the tricks of the trade.

The work presented in this thesis would not have been possible without the financial backing that was provided through an Australian Postgraduate Award and the Grains Research and Development Corporation (GRDC) through a Grains Industry PhD Research Scholarship. I am truly grateful for this. Additionally, I am grateful for the support of the USQ Centre for Sustainable Agricultural Systems that I'm a member of, as well as any other supporting USQ organizational units. I also acknowledge the contribution from the Australian Commonwealth Government through the Research Training Program (RTP) Fees Offset scheme.

To the farm manager, Angus Andrews, and farm agronomist, Scott Ceeney at 'Old Bundama' as part of Hassad (now Paraway Pastoral), I am truly grateful for your assistance and hospitality leading up to, and during the sampling expedition. I am very much appreciative for the lend of sampling equipment provided by Hassad that helped in the efficient extraction of 1200 soil cores over the two week sampling period. The help of Andrew Smart and Michael Wells from Precision

Cropping Solutions in locating an appropriate investigation site and providing an introduction the Angus and Scott has not gone unnoticed.

To Tom Montgomery, I am greatly thankful for the loan of your beasty Holden Rodeo during the sampling exhibition. Despite the jammed Toby Keith CD, which literally remained on repeat for the entire 2 weeks, your vehicle managed to safely navigate shady dirt tracks through the Pilliga Scrub within the wee hours of the morning using erroneous GPS directions. Once surviving that, I knew the remainder of my PhD would be a breeze. Thanks mate.

Thank you to all my fellow students that have helped me greatly over the past few years, both with assistance in the laboratory and office. A special thanks must go to David West who provided much needed assistance in sample preparation and analysis. Westy is a great mate and I look forward to returning the favor during his PhD journey. To Aram, Aaditi and YC, it took some time but I can now definitely conclude your office banter is on-point. Well done!

To Tom, Jane, Tommy and India, you guys have been a massive part of my life prior to, and during this journey, and I am grateful for the massive influence you have had. To my brother Peter, thank you for introducing me to the world of science and engineering from a young age. I have no doubt that without your influence, I would not have developed such an inquisitive mind. To my partner in crime, Lauren, I cannot be thankful enough for your overwhelming support over the last few years. It truly would not have been possible without you, and I am so glad we could journey through it together. Importantly, I would like to extend a big thanks to my Mother, and number one fan, Merilyn Roberton (Donny), who has provided me with a quality education and an appreciation for hard work and determination. I will be forever grateful.

Table of contents

List of tables

List of figures

Abbreviations

AAS	Atomic adsorption spectrophotometer	NDVI	Normalised difference vegetation index
ANN	Artificial neural network	NLML	Non-linear machine learning
BD	Bulk density	NSW	New South Whales
BMP	Best management practices	OK	Ordinary kriging
BNN	Bayesian belief network	PA	Precision agriculture
BR	Blanket-rate	PC	Principal components
CEC	Cation exchange capacity	PCA	Principal component analysis
cLHS	Conditioned Latin Hypercube sampling	PMP	Permanent monitoring points
CPT	Conditional probability table	PSS	Proximal soil sensing
CTF	Controlled traffic farming	PTF	Pedotransfer function
DSM	Digital soil mapping	RMSE	Root-mean square error
DST	Decision support tools	ROI	Return on investment
EC	Electrical conductivity (directly proportional to electrolyte concentration)	RTF	Random traffic farming
ECa	Apparent electrical conductivity	RTK	Real-time kinematic
EM	Electromagnetic induction	SCANS	Soil Condition Analysis System
ESP	Exchangeable sodium percentage	SD	Standard deviation
GPS	Global positioning system	SSCM	Site-specific crop management
GR	Gypsum requirement	SSPF	Soil spatial prediction function
ID	Internal diameter	SSPFE	Soil spatial prediction function with autocorrellated errors
ML	Machine learning	SVM	Support vector machine
MLR	Mixed-linear regression	VR	Variable-rate

1. General Introduction, aims and thesis overview

1.1. Introduction

Australian agricultural producers do not value data to the level required for true precision management. This is due to the perceived cost of data acquisition and limited understanding of the managerial benefits that arise from such an investment (Bennett and Cattle, 2014; Lobry de Bruyn and Andrews, 2016). Subsequently, many critical on-farm decisions are based on limited data, with no real indication of the likelihood of a desired result. This is particularly true for soil constraint management, where large economic investments are committed over a highly spatially variable soil resource, with minimal to no soil data. Farming records pertaining to seasonal change, operational systems etc. provide useful year-to-year data on which to base standard operation procedures for changing circumstance at the paddock scale, but could be enhanced through the collection and utilisation of greater farming-system data to allow true precision agricultural advice within paddock scale. Therefore, achieving true precision management requires a greater volume of farm-specific data coupled with advanced analytical and reporting mechanisms that support decisions (Kelly et al., 2017).

One of the largest limitations of conventional agriculture is the severe lack of depth-specific soil information available at the farm-level (Lobry de Bruyn and Andrews, 2016; McKenzie et al., 2003), which in some instances, is virtually non-existent. Whilst agriculture is advancing into an age where on-farm data collection is increasing (daily satellite imagery, annual crop yield mapping, farm-specific weather data etc.), the majority of data streams rely on surface, or near-surface remote sensing in the x, y spatial plane (Atzberger, 2013). Some technologies aim to capture soil information to depth (electromagnetic induction, Gamma-Ray Spectrometry, NDVI imagery), however, they are unable to measure or detect specific soil properties at an exact depth, and instead provide a depth-weighted integration of soil factors. A better approach towards understanding soil function to depth involves the combination of these methods, and other technologies that may exist, using a data fusion approach.

There are three approaches for capturing soil data, namely: i) direct measurement (via sampling laboratory analysis); ii) remote sensing; and, iii) proximal soil sensing (PSS). Whilst

direct measurement provides us with the most accurate result, it is highly laborious and costly in nature, therefore inhibiting the feasibility of capturing data at a high resolution (both spatially and temporally) (Viscarra Rossel and Bouma, 2016). As such, conventional sampling methods are often applied at a sparse resolution and do not accurately represent the variation which is present within a given field (Viscarra Rossel et al., 2011). To simply take more data using current sampling approaches is not necessarily feasible, especially where laboratory analysis costs are prohibitive. Knowing which data to collect could limit the suite of analyses, but the literature detailing soil health/quality indicators is numerous, with varied indicators and no strong agreement of key criteria (Bennett and Cattle, 2013; Bennett and Cattle, 2014; Bünemann et al., 2018). Furthermore, there is a need for a culture that values data collection and the use of data to inform decision making processes. Within agriculture this is usually limited (Lobry de Bruyn et al., 2017), and is somewhat of a circular argument: More data is needed, but will only be taken where data is valued, while value in data only comes from taking more data and observing the positive results. Thus, there is requirement to motivate farming enterprises to take more meaningful data that will lead to prescriptive decisions, and the role of indirect measurement will be vital in driving this initially. The quantum of data required will need to be demonstrated in making the business case for further data collection (McBratney et al., 2003; McBratney et al., 1981).

Remote sensing provides some capability for increasing the resolution of data capture, at least at the land-surface. Common remote sensing methods currently adopted in the field of soil science stem from optical sensors mounted on either satellites (e.g. Landsat, Sentinel, Rapid Eye, Dove etc.) or drones that capture different bands of surface reflectance (e.g. colour, infrared, hyper spectral, microwave etc.) (Zribi et al., 2011). Such bands may be used to directly estimate soil surface conditions, such as soil moisture (Lakshmi et al 2013), soil texture, soil organic carbon (Gomez et al., 2008), soil salinity (Metternicht and Zinck, 2003) and topography, or indirectly estimate soil conditions via estimating vegetation cover and type. Spatial resolution has previously been an issue with this technology, however recent advances have addressed this (see ‘Dove constellation’, Planet Labs Ltd). Obtaining such information may provide an indication of some general within-field trends, and certainly provides a rapid means of collecting information. However, the inference is still constrained to the land-surface and to the soil parameters that can be inferred from reflectance. In order to enhance this technology’s usefulness, in terms of soil

condition at depth, an integration of other field-based data streams with greater spatial density is needed.

PSS was effectively introduced to augment remote sensing and deliver spatial data streams which were rapid and inexpensive to collect in comparison to direct measurement (Viscarra Rossel et al., 2010a). The PSS approach has driven the crux of precision agricultural development in recent history (McBratney et al., 2003; Viscarra Rossel and Bouma, 2016), becoming a popular method for obtaining soil information, due to recent technological advancements in sensors. Many methods for PSS exist that are used to infer or directly measure an array of soil attributes using a range of techniques and covering a wide range of the electromagnetic spectrum from ultraviolet, visible and infrared light, radio waves, gamma, x-ray and neutron (Rossel et al., 2011). PSS technology has traditionally been limited to detecting surface conditions, or subsurface conditions as a depth-weighted integration only, without the capability to directly target responses at a specific depth. However, recent advances have aided in the collection of proximally sensed data to depth using spectral responses to soil attributes in the visible-near infrared (vis—NIR) range. Measurement to depth is performed using penetrometer based systems (e.g. Veris) or on extracted soil cores combined with other techniques such as gamma attenuation for the direct measurement of bulk density (BD) (see Soil Condition Analysis System, SCANS, Viscarra Rossel et al. 2017).

Beyond research, the adoption of such technology is largely limited at this point in time. This is both due to the technology still being at a juvenile stage requiring expertise in the development and calibration of spectroscopy models, availability of spectral libraries to support calibration, and the initial capital outlay of spectroscopic sensors (e.g. approximately 100K AUD for a visible-near infrared spectrometer). However, recent technical developments e.g. in microelectromechanical systems (MEMS) spectrometers may address these cost limitations. Another limitation is that to measure many soil properties with vis-NIR spectroscopy requires the development of empirical calibrations using laboratory measurement on local calibration samples. The quantum of this data will depend on the scale of soil survey and the inherent variability of the landscape. Regardless, systems like SCANS can provide a significant improvement in the capability to assess spatial (3-dimensional) and temporal soil variability by combining PSS technology with other data layers (e.g. yield maps or remote sensing) and geostatistical methods (Viscarra Rossel et al., 2017).

The deployment of PSS technology requires significant understanding of its operating mechanisms, limitations, data inference, and data manipulation, which subsequently demands specialised technical skills. The PSS technology that is commercially available and readily deployable by operators (such as EMI) can only measure bulk soil properties (e.g. electrical conductivity [EC]). Hence, while PSS augments the capability to infer soil spatial and temporal information in the x, y, z planes, there is still a requirement for direct data (calibration) and a highly skilled human resource to interpret these sensor measurements and provide agronomically useful soil information.

One way this can be achieved is to generate secondary inference of soil properties via the application of pedotransfer functions (PTFs) and soil spatial prediction functions (SSPF) (McBratney et al. 2018) which aim to predict unknown soil properties from available soil and environmental data (McBratney et al., 2003; McBratney et al., 2002). Therefore, PSS in conjunction with PTFs and SSPFs provide the ability to develop Digital Soil Maps (DSMs) of key soil properties, facilitating a basis for precision management.

PTFs have often been developed using a spatially diverse dataset, thus usually resulting in a large prediction error. Such error may be acceptable for detecting differences at a regional level, but is considered too large to detect differences at a sub-paddock level; i.e. the spatial resolution of training data is too coarse to meaningfully design within field variable rate management that is linked to yield variability and the interacting soil constraints affecting this. Whilst previous sampling methodologies based on geostatistical analysis have been developed to provide a localised calibration dataset for PTFs (Brus and De Gruijter, 1997; Brus et al., 2004; Kennard and Stone, 1969; McKenzie and Ryan, 1999; Minasny and McBratney, 2010; Minasny and McBratney, 2006b), there is a limited understanding surrounding the size of the dataset required to achieve acceptable spatial predictions for sub-paddock precision management. Furthermore, the vast majority of PTFs are purely developed on empirical models which are incapable of incorporating qualitative management information. Management information, and decisions, are well known to have a large degree of influence on soil function (e.g. compaction and random in-field traffic versus spatially controlled traffic; Bartimote et al., 2017; Bennett et al., 2017). This data is difficult to capture, especially in a volume and replication level appropriate for inclusion in linear regression and nonlinear approaches which are required to explain the complexity of farming systems

(Bennett et al., 2019). On this basis an approach is required that incorporates the system complexity, including qualitative and semi-quantitative management data, where empiricism may not be the most suitable approach. The argument can be made that probabilistic approaches (McBratney et al., 1981), and even artificial intelligence approaches, are suitable to aid in the determination of system variability and subsequent management. Of course, this requires judicious application, and in most cases a measure of uncertainty would be needed, which could preclude artificial intelligence for broader application. This will be an ongoing global scientific discussion, and will be considered throughout this thesis.

Together PTFs and SSPFs provide a very powerful capability for understanding soil variability, although a framework for their employment at the practitioner level is largely missing. This is not a failure of the fields, but an indicator of the fact that these fields are still in developmental stages. That said, there are certainly aspects of both PSS and PTF which could have been better exploited in terms of commercial precision agriculture. Throughout this thesis the framework for practical implementation of PSS and PTFs will be considered, with a synthesis of this provided in the general discussion.

Even though PSS and PTFs currently provide the greatest opportunity to satisfy data requirements for precision management, they are based on the assumption that agricultural practitioners do not value data as a tool for on-farm application or management. In Australia, and many other countries, this has certainly been a correct assumption, consequently leading to the reliance on PSS and PTF use.

The agronomic applications of these techniques have been limited to quantifying general soil variability to better inform management zones or strategic location of directly measured soil information (sampling and laboratory analysis) to maximise the value of the soil survey resource (McBratney et al., 2003). This thesis seeks to challenge the status quo of data-value by demonstrating that significant investment in directly measured soil data to depth, describing the structural nature of soils, will greatly enhance the diagnostic and predictive capabilities at the localised scale. The focus of this work is to ascertain the quantum of data required on-farm against the inherent variability, and crop production metrics, whereby the output of the work is to demonstrate clear value to invest in soil data. This will inform a framework for powerful localised calibration datasets for soil constraint management. The complex and high dimensional nature of

such calibration datasets (i.e. large number of variables for each observation) and the potential size in which they may exist lends themselves to investigate other mathematical approaches. These approaches may be more desirable for PTF development and calibration rather than traditional methods such as linear regression. Arising from this investigation are some initial key questions which this thesis seeks to address:

1. How much direct data is required to formulate a meaningful soil property – crop response calibration dataset that can inform on-farm variable rate management?
2. How does sampling design affect the accuracy of agronomic advice?
3. How can qualitative data be included in on-farm management spatial management?
4. Assuming the above questions can be answered, what is the implementable framework and who is the service provider?

1.2. Aims – data needs for amelioration

Limited attention has been paid to the consequences of a data minimal approach to agricultural soil constraint management. Consequently, soil amelioration advice is often spatially inaccurate, causing the potential for large economic losses. Precise management of constraints requires a thorough understanding of soil properties and their spatial dependence at the sub-paddock scale, and knowledge on site-specific constrain-yield interactions to guide amelioration investment advice based on potential return on investment (ROI). Whilst PSS and RS provide useful data streams to characterise soil variability, they fail to accurately measure and define the soil properties which describe site-specific constraints and their interactions, especially in the sub-surface soil layers. Therefore, soil constraint management requires significant laboratory measurements to guide amelioration advice.

Current DSM approaches offers the ability to provide spatial soil amelioration advice. However, they are inhibited by the lack of directly measured soil data and, therefore, cannot represent site variability to the level required for true precision management. This results in the development of variable rate prescriptions for soil amendments that are not well suited to site conditions. Improving the economic and agronomic efficiencies of site-specific recommendations requires consideration of the cost of data acquisition, the cost of spatial recommendation errors and the economic benefit of a yield response due to an amendment application. This will be explored by using a spatially exhaustive dataset to firstly investigate the site-specific spatial

predictions errors associated with current DSM approaches at varied sampling densities; and secondly by applying machine learning (ML) techniques to identify soil-crop interactions.

It follows that the study has these four main aims:

- Demonstrate the agronomic consequences of a data minimal approach to soil sampling
- Explore the accuracy of digital soil mapping (DSM) approaches to spatially predict soil properties
- Investigate the merit of linear and nonlinear ML approaches to reveal key site-specific soil-crop interactions
- Explore the ability to capture and integrate qualitative management information to inform decision making

1.3. Thesis overview

1.3.1. Chapter 1: General introduction, aims and thesis overview

This chapter introduces the broad issues associated with soil data collection and current spatial soil management practices in precision agriculture. It identifies various questions surrounding spatial data collection and application, and specifies the aims of the study in relation to these.

1.3.2. Chapter 2: An investigation into the current and future trends of machine learning for spatial management of soil information: A review

This chapter presents background information surrounding the use of data for spatial soil management in agriculture. This chapter includes: an investigation of spatial data streams currently available in precision agriculture; a brief overview of DSM techniques and their application; and, an investigation of four key ML approaches, identifying key areas of application in the literature and highlighting opportunities for future adoption.

1.3.3. Chapter 3: Description of the experimental site and soil analytical methods employed in this work

This chapter describes the investigation site used in this study, providing detailed information pertaining to the geographical location, climate, history of the site, spatial auxiliary data available for the site and the variability of soil structural components measured at the site. A

detailed description of the sampling techniques and methods used for the analysis of soil structural components is also presented.

1.3.4. Chapter 4: Assessing the sensitivity of site-specific lime and gypsum recommendations to soil sampling techniques and spatial density of data collection: A pedometric approach

This chapter investigates the error of site characterisation using four sampling and spatial prediction techniques, namely; i) bulked transect sampling; ii) spatial clustering; iii) ordinary kriging; and, iv) regression kriging. This is assessed by simulating the application of each technique at sampling densities of 10, 20, 50, 100, 150, 200, 250 and 300 samples/108 ha using the directly measured spatial dataset collected for this body of work. The magnitude of error for each technique is quantified using the root-mean-square-error (RMSE) of predictions of ESP and pH, and subsequently against the error of recommendation advice for variable-rate application of gypsum and lime for soil amelioration. For this chapter, recommendation error is quantified by the magnitude of over- and under-application of each ameliorant. The magnitude of error associated with random initialisation of each technique is also investigated.

1.3.5. Chapter 5: Crop yield prediction using machine learning for the purpose of soil constraint diagnosis

This chapter seeks to develop linear and nonlinear site-specific yield prediction models for a single site using the spatially intensive soil structural dataset collected for this body of work. The models investigated were; i) mixed-linear regression; ii) Cubist; iii) artificial neural network; and, iv) support vector machine. For each technique, single-season models were trained and validated to the spatial yield data available for 3 individual wheat cropping seasons (i.e. 2013, 2015 and 2016). Attempts are also made to develop a temporally stable yield prediction model for the site using multiple years of yield and weather data. The effects of data size and data dimensionality on model generalisation are investigated, as well as an attempt to assess model performance and the presence of overfitting beyond the R^2 metric. The requirement for improved interpretability methods to assess generalisation of ML approaches are further discussed.

1.3.6. Chapter 6: Towards identifying the soil data investment to economically optimise soil ameliorant recommendations as a function of yield

This chapter aims to identify the minimum dataset required to economically optimise the application of gypsum as a soil ameliorant for the investigated site, using a variable rate approach. This is achieved by applying spatial prediction methods from Chapter 5, and the Cubist yield model adapted from Chapter 6, to estimate the economic errors associated with the under- and over-application of gypsum at various sampling densities, while further providing consideration toward the cost of sampling. Using the developed yield prediction model to estimate yield response due to application, the long-term economic benefits of soil amelioration are considered. This chapter further discusses the requirement for a variable rate approach in soil constraint management, as well as highlights the necessity for increased soil data investment in precision agriculture.

1.3.7. Chapter 7: A Bayesian approach toward the use of qualitative information to inform on-farm decision making: The example of soil compaction

Empirical approaches fail to capture qualitative information which may be highly influential on system dynamics and subsequently the decision-making process. Therefore, this chapter seeks to investigate a Bayesian belief network (BBN) approach to merge quantitative information with qualitative relationships to inform the risk of soil compaction, as an example. The applied network utilises soil moisture information, soil characterisation information, soil loading conditions and management information to infer the level of soil compaction risk associated with any given field operation. Consideration is provided to further develop the applied BBN to serve as a decision-support tool by utilising soil stress models and soil moisture models. The application of qualitative approaches to spatial soil management is further discussed.

1.3.8. Chapter 8: General discussions, conclusions, and future research

This chapter more broadly discusses the key findings in each chapter presented in this work, highlighting their significance toward improved spatial soil management in precision agriculture. The required change in perceptions pertaining to data value in Australian agriculture is considered, as well as a discussion pertaining to the future development of a conceptual framework that incorporates multiple data sources and empirical techniques to better inform spatial management.

Further consideration of the service provider required to implement these approaches is also discussed.

1.4. References

Atzberger, C., 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote Sensing 5(2), 949-981.

Bartimote, T., Quigley, R., Bennett, J.M., Hall, J., Brodrick, R., Tan, D.K., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: II. Economic and physiological analysis. Soil and Tillage Research 168, 133-142.

Bennett, J.M., Cattle, S., 2013. Adoption of soil health improvement strategies by Australian farmers: I. Attitudes, management and extension implications. The Journal of Agricultural Education and Extension 19(4), 407-426.

Bennett, J.M., Cattle, S., 2014. Adoption of soil health improvement strategies by Australian farmers: II. Impediments and incentives. The Journal of Agricultural Education and Extension 20(1), 107-131.

Bennett, J.M., Roberton, S., Marchuk, S., Woodhouse, N., Antille, D., Jensen, T., Keller, T., 2019. The soil structural cost of traffic from heavy machinery in Vertisols. Soil and Tillage Research 185, 85-93.

Bennett, J.M., Roberton, S.D., Jensen, T.A., Antille, D.L., Hall, J., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: I. Heavy machinery impact on the soil resource. Soil and Tillage Research 168, 143-154.

Brus, D., De Gruijter, J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80(1-2), 1-44.

Brus, D., De Gruijter, J., Van Groenigen, J., 2004. Designing purposive and random spatial coverage samples by the k-means clustering algorithm, Global Workshop on Digital Soil Mapping, Montpellier, France.

Bünemann, E.K., Bongiorno, G., Bai, Z., Creamer, R.E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T.W., Mäder, P., 2018. Soil quality–A critical review. Soil Biology and Biochemistry 120, 105-125.

Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146(3-4), 403-411.

Kelly, N., Bennett, J.M., Starasts, A., 2017. Networked learning for agricultural extension: a framework for analysis and two cases. The Journal of Agricultural Education and Extension 23(5), 399-414.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11(1), 137-148.

Lobry de Bruyn, L., Andrews, S., 2016. Are Australian and United States farmers using soil information for soil health management? Sustainability 8(4), 304.

Lobry de Bruyn, L., Jenkins, A., Samson-Liebig, S., 2017. Lessons learnt: sharing soil knowledge to improve land management and sustainable soil use. Soil Science Society of America Journal 81(3), 427-438.

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

McBratney, A., Webster, R., Burgess, T., 1981. The design of optimal sampling schemes for local estimation and mapping of of regionalized variables—I: Theory and method. Computers & Geosciences 7(4), 331-334.

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109(1), 41-73.

McKenzie, N., Bramley, R., Farmer, T., Janik, L., Murray, W., Smith, C., McLaughlin, M., 2003. Rapid soil measurement–a review of potential benefits and opportunities for the Australian grains industry. GRDC Project CSO 27.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89(1-2), 67-94.

Metternicht, G., Zinck, J., 2003. Remote sensing of soil salinity: potentials and constraints. Remote sensing of Environment 85(1), 1-20.

Minasny, B., McBratney, A., 2010. Conditioned Latin hypercube sampling for calibrating soil sensor data to soil properties, Proximal soil sensing. Springer, pp. 111-119.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & geosciences 32(9), 1378-1388.

Rossel, R.V., Adamchuk, V., Sudduth, K., McKenzie, N., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time, Advances in agronomy. Elsevier, pp. 243-291.

Viscarra Rossel, R.A., Adamchuk, V., Sudduth, K., McKenzie, N., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time, Advances in agronomy. Elsevier, pp. 243-291.

Viscarra Rossel, R.A., Bouma, J., 2016. Soil sensing: A new paradigm for agriculture. Agricultural Systems 148, 71-74.

Viscarra Rossel, R.A., Lobsey, C.R., Sharman, C., Flick, P., McLachlan, G., 2017. Novel Proximal Sensing for Monitoring Soil Organic C Stocks and Condition. Environmental Science & Technology 51(10), 5630-5641.

Viscarra Rossel, R.A., McBratney, A., Minasny, B., 2010. Proximal soil sensing.

Zribi, M., Baghdadi, N., Nolin, M., 2011. Remote sensing of soil. Applied and Environmental Soil Science 2011.

2. An investigation into the current and future trends of machine learning for spatial management of soil information: A review

2.1. Introduction

The farming system is inherently variable and difficult to manage. The philosophy of 'precision agriculture' (PA) aims to account for this variability by adjusting management strategies and farming resources to appropriately match these conditions, with the intent of optimising overall productivity (Whelan and McBratney, 2000; Zhang et al., 2002). PA is commonly applied in practice via 'site-specific crop management' (SSCM) strategies that aim to spatially and temporally match farming resources to the requirements of soil and crop (Whelan and McBratney, 2000). Whilst weather, disease and pest infestations contribute to crop variability at the sub-field scale (Fiener and Auerswald, 2009; Prabhakar et al., 2012), variation in soil condition is one of the largest contributing factors responsible for spatial variations in crop yield (Whelan and McBratney, 2003). The management of soil at a within-paddock scale is therefore an integral part of the PA philosophy and should remain a focus for improved farm management.

Two sources of error exist that influence the accuracy of SSCM, namely 1) The ability to accurately characterise soil variability; and 2) The ability to diagnose site-specific yield limiting constraints and match the appropriate ameliorant recommendations. Soil conditions are known to be highly variable at small spatial scales (McBratney and Pringle, 1999; Warrick, 2001), which inhibits the ability to accurately characterise soil condition. Empirical and statistical spatial prediction methods exist (McBratney et al., 2003) which aim to map continuous soil properties using a mixture of directly measured, remotely sensed and proximally sensed information. The accuracy of these approaches is largely dependent on the density of directly measured soil information, with limited guidance to suggest ideal sampling densities for an economically optimised agronomic recommendation such as spatially variable: nutrient application, structural amelioration, strategic-tillage/ deep-tillage placement, and seed variety positioning (depth and variety with space). The scale of operation and management interventions ultimately defines the density of data and the accuracy required, although there is little guidance as to the prescribed data density for variable rate management of multiple constraints.

Diagnosing site-specific yield limiting constraints requires knowledge of spatial soil-crop interactions for a given site and can help inform variable rate recommendations for

amelioration. These interactions are known to be highly complex and nonlinear in nature (Dai et al., 2011; Oldfield et al., 2019; Sudduth et al., 1996) and are represented by high-dimensional data (i.e. large number of features). Traditional linear machine learning (ML) approaches, such as linear regression and its linear permutations, are not well suited to this data (Sudduth et al., 1996), due to the structural assumptions they apply which inhibits their ability to map nonlinear trends. Therefore, nonlinear approaches for empirical crop modelling should be investigated.

Understanding spatial soil-crop interactions is not only pertinent in the diagnosis of soil constraints, but is required for the estimation of a yield response due to an induced system change. This allows for economic consideration towards site-specific soil amelioration. This review investigates these approaches for soil spatial prediction before exploring the current and future trends of ML in agriculture. The opportunities and limitations of various ML approaches will be investigated, and their applicability for providing useful insight into spatial soil-crop interactions will be assessed.

2.2. Spatial data stream technology

2.2.1. Machinery

Advances in agricultural machinery offer the capability of sensing and logging on-the-go machinery performance parameters which can be site-referenced using GPS guidance technology (Thomasson et al., 2019). This allows for the mapping of machine performance variables (e.g. fuel use, wheel slip, ground speed and draft force etc.), and machine application variables (e.g. seed rate application, chemical application, depth of tillage etc.). The most widely used machine parameter is that obtained from the yield sensor in harvesting vehicles (Arslan and Colvin, 2002; Pierce et al., 1997). Whilst not related to machine performance, this provides the ability to measure and map crop yield at fine scale, thus providing direct indication toward the spatial variance of soil condition. Even though yield data is an output variable of principal concern to site-specific management, it is an integration of several variables, some of which are temporally dependent. Therefore, multiple years of yield information is required, along with other associated spatial data sources, to diagnose spatial causes of variability pertaining to soil condition.

Machinery performance variables such as fuel use and draft force can be aggregated to develop a soil resistance map to infer spatial changes in soil condition (Sirjacobs et al., 2002; Tsiropoulos et al., 2013; Van Bergeijk and Goense, 1996). Soil resistance cannot predict any

single soil variable, but it may provide useful insight into spatial and temporal soil patterns when leveraged with other spatial data sources, with relative differences potentially able to aid in identification of soil compaction. Advances in machinery technology will consequently increase the volume of on-the-go machinery data, allowing for improved soil inference from these data streams (e.g. soil moisture sensors on tillage equipment). Spatial machinery data such as fuel use provides the ability for economic analysis of any given field operation in the assessment of management decisions (i.e. cost-benefit of deep ripping). Furthermore, recording of historical data allows for spatial interrogation of user or machinery error when an observed result is unexpected (e.g. interrogating applied seed rates and depth over an area of zero emergence). Spatial machinery data is continuously collected during field operations, but it is rarely used to inform management decisions (Pringle et al., 2003). This is due to the inability to interrogate the data in farm management platforms alongside other data sources. Opportunity exists to better utilise this data to inform causes of spatial variability.

2.2.2. Remote sensing

Remote sensing technology has been widely adopted in precision agriculture to identify and map spatial trends by measuring the electromagnetic radiation reflected from a surface (Viscarra Rossel et al., 2011). More recently, sensors have commonly been mounted to drone and satellite platforms, therefore providing the capability for data collection at large spatial and temporal scales. Reflectance can be measured over various wavelength ranges, including visible (Vis), near infrared (NIR), infrared (IR), ultraviolet (UV) and microwave (MW) portions of the spectrum. Application of remote sensing in agriculture has included crop yield prediction (Lobell et al., 2015; Shanahan et al., 2001), crop stress (Barnes et al., 2000; Tilling et al., 2007), quantification of pest and hail damage (Bentley et al., 2002; Genc et al., 2008), soil moisture estimates (Mohanty et al., 2017; Peng et al., 2017) and soil carbon (Mondal et al., 2017). For further information pertaining to the specific use and uptake of remote sensing technologies in PA, readers are referred to the seminal reviews of Mulla (2013) and Atzberger (2013).

Remotely sensed data can be used to assess spatial patterns in crop productivity where yield data is not available, by the calculation of vegetation indices, including, but not limited to NDVI, enhanced vegetation index (EVI), soil-adjusted vegetation index (SAVI) and Red edge. With the ability to access multispectral satellite dating to 1973 (Landsat), this can aid in the identification of temporally stable crop patterns for the purpose of soil constraint diagnosis

(Dang and Moody, 2016). Whilst remotely sensed crop data may provide indication on the level of spatial and temporal variability, the measured reflectance is an integration of soil variables which contribute to plant health, and therefore cannot be used to identify specific soil properties or constraints. Even though methods exist to estimate soil properties (e.g. moisture and carbon), these are constrained to surface conditions, and do not provide insight into depth-based information. A further limitation is that prediction of soil properties spectral reflectance at a remote distance due to a high noise to signal response in comparison to proximal sensing solutions. Therefore, integration of other data sources is required to further leverage remotely sensed information to provide inference of soil condition to depth.

2.2.3. Proximal sensing

Proximal soil sensing (PSS) was adopted to provide a level of accuracy greater than remote sensing, whilst maintaining the ability to capture soil information at large spatial scales. PSS employs the use of sensors that obtain signals from a soil medium at a distance of < 2.0 m (Viscarra Rossel and McBratney, 1998; Viscarra Rossel et al., 2010b).. Proximal sensors are further described by the way measurement is undertaken (invasive or non-invasive), the source of energy used (active or passive), their mode of operation (stationary or mobile) and if the soil property is directly or indirectly measured, as described by (Viscarra Rossel et al., 2011). Spatial mapping of soil properties for PA purposes is primarily concerned with the 'mobile' subset of PSS methods, termed 'on-the-go' PSS. On-the-go PSS provide rapid spatial assessment off soil chemical, physical and mechanical characteristics, albeit from the surface.. Key reviews in this space are given by Adamchuk et al. (2004), Viscarra Rossel et al. (2010b) and Viscarra Rossel et al. (2011).

Electromagnetic and radiometric sensors are the most widely used sensors in terms of commercial purposes within PA (Adamchuk et al., 2004). Whilst these sensors are used to estimate soil characteristics directly (Van Egmond et al., 2010), they are more commonly used as environmental covariates for indirect measurement, via application of a soil-spatial prediction function (SSPF) combined with site-based calibration soil core samples (Minasny and McBratney, 2010; Wong et al., 2010). On-the-go PSS is limited though in its ability to capture depth-specific information, as surface measurements typically only provide soil factor information as a single-value, depth-weighted integration. Furthermore, the range of soil characteristics which correlate well with measurements obtained from on-the-go proximal sensors is highly limited.

Recognising the limitations of on-the-go PSS to define spatial variability of soil characteristics, recent advancements in technology have allowed for depth-based measurements combining both on-the-go PSS and stationary PSS. Therefore, the number of soil characteristics that can be estimated is increased vastly. In-situ depth-based estimates of soil properties have been achieved using research-scale equipment (Baharom et al., 2015; Kusumo, 2018; Kusumo et al., 2008; Waiser et al., 2007) , which are largely limited by the speed of measurement. Veris technologies (Veris Technologies, Salina, KS, USA) offer some commercial solutions for on-the-go surface and shallow subsurface (0–7.6 cm) vis-NIR measurements, as well as in-situ depth-based measurements. Whilst their in-situ platform (P4000), allows depth measurements to 137 cm, it is constrained to within the 397–2212 nm (vis-NIR) spectral range (Zhang et al., 2017), and excludes important absorption features diagnostic of possible soil properties that can be estimated, e.g. phyllosilicates in the 2200–2800 nm range.

The 'soil condition analysis system', SCANS, (Viscarra Rossel et al. 2017) is an example of combining field-based vis-NIR and electromagnetic sensing equipment with an integrated coring system to obtain spectroscopy measurements at a fine depth resolution along an extracted core. The addition of three spectroscopic sensors enable quantitative and rapid estimates of bulk density (BD), organic carbon content and composition, soil water, texture and cation exchange capacity (CEC) to a depth of 1.2 m (Viscarra Rossel et al. 2017). Whilst the SCANS system may overcome many limitations of PSS, the technology platform is still in evaluation states. The field deployment of SCANS is currently limited by operational costs as it still relies on human operations, although could be developed as a fully automated driverless implement. The coupling of SCANS with predictive soil mapping techniques to interpolate the measured points from this platform to a finer spatial resolution require further evaluation.

2.2.4. Spatial data stream technology fusion

Benefit exists to fuse spatial data steams into a single system to provide useful soil insights for variable soil management. However, at the current level of soil data collection, this is not currently possible at the level required. The use of on-the-go and static PSS systems provides a pathway, but localised calibrations are still required, and the true power of this approach cannott be realised until a critical mass of data exists. This should not detract from the technology, but reinforce the requirement to better understand the sheer volume of data that is needed, within the constraints of time and capital. Adoption of this technology is conditional

on the development of a commercial market , which will necessitate the data outputs to be demonstrated as useful in informing agronomic recommendations (Bennett and Cattle, 2013; Bennett and Cattle, 2014; Lobry de Bruyn, 2019).

Furthermore, while agricultural machine generated data and remotely sensed data provide valuable data streams, the approximate 20 years of (Bishop and McBratney, 2001; Boydell and McBratney, 2002; Chang and Islam, 2000; Moran et al., 1997; Stafford et al., 1996) has not achieved reliable diagnosis of crop yield constraint mechanisms, or predict subsequent yield reliably. This suggests that the problem of yield variability is highly complex, and that use of the plant as an integrated sensor of soil depth information may not be appropriate. On this basis, it is purported that direct measurement soil depth characteristics is required to augment the fusion of spatial data stream technology networks, but that this will also be reliant on the ability to geostatistically fuse the data streams.

2.3. Digital soil mapping (DSM)

2.3.1. Geostatistical and interpolation approaches

Geostatistical and interpolation approaches to soil mapping are based on the below formulation (Equation 2.1.), where soil at some location (x,y) is dependent on the geographic coordinates (x,y) and soil at some neighbouring location (x+u, y+v) (McBratney et al., 2003).

$$S = f(x, y), s(x + u, y + v) \qquad \text{Equation 2.1.}$$

Geostatistical and interpolation approaches aim to spatially interpolate soil information to a finer resolution than it is collected/ available at, by application of various mathematical techniques. A number of approaches exist, namely:

i) Inverse squared distance interpolation (Laslett et al., 1987);
ii) Natural neighbour interpolation (Sibson, 1981);
iii) Quadratic trend surface (Lark and Webster, 2006);
iv) Laplacian smoothing splines (Wendelberger, 1981); and,
v) Ordinary kriging (OK) (Burgess and Webster, 1980a).

Of these, OK has been the most widely adopted interpolation technique in soil science (McBratney et al., 2000). Interpolation is based on fitting a semi-variogram to the observation data in order to describe the relationship between distance and variance of a given soil property (Figure 2.1.). Common variogram models can be linear, spherical, exponential or Gaussian

functions, depending on the natural variance presented in the observational data and are fitted to the data using a weighted least squares method (Cressie, 1985; McBratney and Webster, 1986). All models should be tested to ensure an appropriate variogram fit to the presented data. The model type that best represents the data is that which has the smallest residual sum of squares or smallest mean square error (Oliver and Webster, 2014).

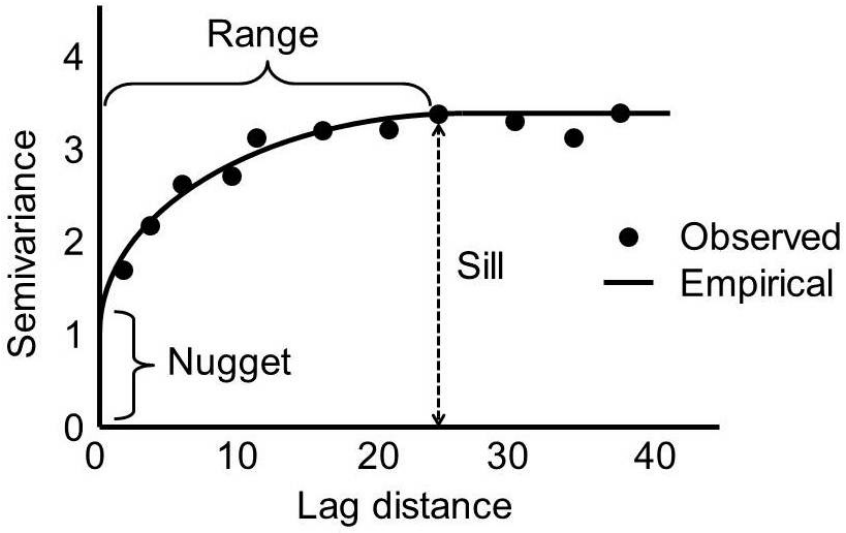


Figure 2.1. Example of semi-variogram fitted to observational data. Nugget, range and sill are parameters of the variogram model, where the nugget represents short range variability or error in the data, and the sill and range represent the semivariance and distance at which spatial autocorrelation is no longer present.

OK approaches are constrained by the quantity of data required to obtain a reliable model of the variogram. Webster and Oliver (1992) suggested a minimum of 100 points are required to satisfy an appropriate variogram fit representing the spatial autocorrelation. Even though decreasing sample density may significantly increase error of the variogram model, the model for practical purposes may still hold relevance for areas with sparse available spatial data. As sampling and analysis can be costly, this should be considered in the design and agronomic cost of spatial prediction error when determining an optimal sampling density. Recognising the limitations of the perceived data requirements for practical application in PA, the adoption of OK has been focused on augmenting statistical DSM approaches through spatial interpolation of model errors. This is termed regression kriging, and is explored further in the following section.

2.3.2. Non-interpolation approaches

The data availability of directly measured soil characteristics is generally sparse, in comparison to environmental covariates which may be collected at a finer spatial resolution. Recognising this, McBratney et al. (2003) extended on the work of Jenny (1941) to formalise the DSM framework called *scorpan* (Equation 2.2.). The framework is based on the underlying assumption of environmental correlation between soil properties and environmental covariates that can be acquired at a greater spatial resolution and reduced cost. Scorpan models are also referred to as soil spatial prediction functions (SSPF; Malone et al., 2018). The aim is to fit numerical models between soil observations and scorpan factors (Minasny and McBratney, 2016):

$$S_a = f(s, c, o, r, p, a, n) + e \qquad \text{Equation 2.2.}$$

which assumes any soil property, '*a*', at a given location is a function of other soil properties at that location based on: space (*s*), climate (*c*), organisms (*o*), relief (*r*), parent material (*p*), age (*a*), and auto-correlated errors (*e*). The addition of auto-correlated errors allows for a hybrid geostatistical non-interpolation approach. The residuals of the developed model are spatially propagated using an OK approach, providing proven superior results to straight geostatistical or non-interpolation approaches (Hudson and Wackernagel, 1994; Knotters et al., 1995; Odeh and McBratney, 2000). This is referred to in the literature as regression kriging (Knotters et al., 1995; McBratney et al., 2003; Odeh et al., 1995) or SSPFe (Malone et al., 2018; McBratney et al., 2003).

SSPFe commonly rely on remotely sensed, proximally sensed and machine generated data (i.e. crop yield data) as environmental covariates for spatial predictions (Dlugoß et al., 2010; Malone et al., 2018; Minasny and McBratney, 2010; Minasny and McBratney, 2007; Niang et al., 2014). However, not all soil properties correlate well with proximally and remotely sensed covariates, meaning additional information is often required to perform predictions. A further limitation of these data streams is that their response is largely driven by surface characteristics that may occur as overtones or incorrectly represent subsoil processes and properties, leading to higher uncertainty in spatial prediction of below surface conditions (Odgers et al., 2015).

Many mathematical models are used to represent the developed SSPF. These include, but are not limited to linear regression (Florinsky et al., 2002; McKenzie and Ryan, 1999; Minasny and McBratney, 2010), regression-trees (Henderson et al., 2005; Lacoste et al., 2014;

McKenzie and Ryan, 1999), support vector machines (SVMs) (Ballabio, 2009; Were et al., 2015) and artificial neural networks (ANNs) (Behrens et al., 2005; Chang and Islam, 2000; Dai et al., 2011). Considerations toward data availability, data dimensionality and the complexity of the relationships between variables are required for model selection to ensure the complexity of the model is matched to the given problem. This helps ensure the models are appropriately fitted to the data and can represent the general trends presented.

Non-interpolation approaches use remotely and proximally sensed information to supplement the intensive directly measured data requirements of geostatistical methods. This is a key advantage of soil mapping for precision agriculture, which is constrained by the cost of sampling and analysis. Whilst these spatial prediction methods provide useful information to the practitioner, there is a requirement for this to be leveraged to inform decision making. Without this crucial step, the outputs of SSPFs become powerless. Therefore, data mining approaches that provide insight into soil spatial predictions and their relationship with production metrics should be investigated.

2.4. Machine learning for spatial soil-crop relationships

Advancements in sensing technologies and DSM techniques have increased the spatial data resolution within conventional agriculture. This data only holds value for agricultural purposes when used in data-driven approaches to provide useful insight into system dynamics for improved management. ML as an automated model building process applies numerous mathematical techniques aimed at identifying complex patterns and relationships within data (Bishop, 2006; Witten et al., 2016). ML tasks are undertaken using supervised or unsupervised techniques, depending on what level of prior knowledge (or data) is available as samples for the inference method. Supervised learning techniques are used to find key relationships between a set of input and output variables from a given set of observations (Kubat, 2015). This is referred to as model training, or learning, and can be achieved for regression problems (i.e. continuous outputs) or classification problems (i.e. discrete outputs) (Kotsiantis et al., 2007). Unsupervised techniques however aim to detect the underlying structure in each dataset, by separating the data into discrete subsets based on feature similarity. This is commonly referred to as data clustering.

Linear ML approaches have traditionally been applied in data limited environments where the effects of outliers and noise can be more influential on model convergence (i.e. the ability for a model to find the correct global solution). These methods, commonly referred to

as high-bias low-variance models, assume a linear structure within the data, therefore preventing model overfitting at the risk of oversimplifying the relationships (Behmann et al., 2015). Nonlinear ML methods however do not assume a data structure, and instead aim to locate the underlying structure within the data. Whilst this allows more complex nonlinear patterns to be identified, the absence of structural assumptions lends the applied models to be susceptible to overfitting to the data, therefore failing to generalise to the given problem (Wani et al., 2019). The data density at which nonlinear ML methods become superior over linear ML methods is inherent to the applied modelling problem, and is dependent on data complexity. Therefore, nonlinear methods should be directly compared to linear methods for a given problem, to ensure the application is appropriate for the more advanced method.

The risk of overfitting and poor generalisation increases as data density decreases, where the effects of outliers and noise become more influential. Generalisation is commonly assessed using cross-validation techniques, where the original data is partitioned into training and validation subsets (Hawkins et al., 2003), and the quality of fit between the two datasets is compared. A model may be considered generalised if the quality of fit between training and validation is of a small magnitude (Wani et al., 2019). Widely used cross-validation techniques include the hold-out method (Kohavi, 1995), where the data is randomly partitioned once, and k-fold cross validation (Efron and Tibshirani, 1994), where random partitioning and model training is repeated. K-fold cross validation allows for more robust and efficient validation, as the prediction variability due to random partitioning can be assessed (Brus et al., 2011; Kohavi, 1995).

Opportunity exists to investigate the use of nonlinear ML approaches in providing inference towards soil-crop interactions which are known to be highly complex and nonlinear (Drummond et al., 1998; Irmak et al., 2006; Sudduth et al., 1996). Four widely adopted nonlinear ML techniques are further investigated to assess the potential in the context of spatial management in precision agriculture. Two of these techniques are focused on regression based problems, namely, i) artificial neural networks (ANN), and ii) support vector machines (SVM). Data clustering will be investigated as a third group of techniques as well as Bayesian belief networks (BBN) in instances where qualitative data is present. The mathematical theory behind each method will be presented, as well as a brief review on key areas of application within the context of agriculture.

2.4.1. Artificial neural networks

2.4.1.A. Theory of artificial neural networks

ANNs are a well-recognised and widely adopted supervised ML technique which can be used for both classification and regression. A strength of ANNs is their ability to model complex and nonlinear relationships between variables (White, 1989) making them suitable for modelling many soil-crop interactions. Whilst an array of ANN architectures exist, the widely used feed-forward back-propagation structure has proven abilities for regression in nonlinear environments (Haykin and Network, 2004) and is presented here. ANNs use a set of connection weights, *W,* and biases, *θ,* as parameters to learn relationships between input and output variables though a multi-stage process, depending on the number of layers used in the network. Network predictions are made by computing the transfer function (Equation 2.3.), followed by the activation function (Equation 2.4.) at each node, *j* within each layer, *l* of the network equation at every node within the network, in conjunction with the associated activation function. Whilst a number of activation functions exist, the most widely adopted is the sigmoid function (Buhmann, 2003; Widrow and Lehr, 1990), due to its superior computational efficiency whilst maintaining predictive performance (DasGupta and Schnitger, 1993). The sigmoid activation function is therefore presented here.

$$O_j^l = \sum_{i=1}^{n} \left(W_{ij} . x_i^{l-1} \right) + b_j^i \qquad \text{Equation 2.3.}$$

Where O_j^l = transfer function at jth node in the lth layer, W_n^j = connection weight of the ith observation and jth node, x_i^{l-1} = output value of the ith node from the (l-1) layer and b_j^i= bias of the jth node at the lth layer.

$$f\left(O_j^l\right) = \frac{1}{1 + e^{(-O_j^l)}} \qquad \text{Equation 2.4.}$$

Network training involves the optimisation of this set of weights and biases such that the error, or cost of the model is minimised (i.e. difference between predicted and observed values is minimised). This is an iterative process whereby the training algorithm back-

propagates errors through the network and minimises the cost function by adjusting the network weights and biases. The most widely adopted training algorithms to optimise the network to a set of observations include gradient descent, Levenberg Marquardt, conjugate gradient and Bayesian regularisation.

2.4.1.B. Utilisation of artificial neural networks in agriculture

ANNs have been well utilised in the field of agriculture (Table 2.1. and Table 2.2), with advancements being focused into two distinct areas of research, namely:

1) Pedometrics – (Baker and Ellison, 2008; Hopmans et al., 2003; Koekkoek and Booltink, 1999; Merdun et al., 2006; Minasny et al., 2004; Minasny and McBratney, 2002; Minasny et al., 1999; Pachepsky et al., 1996; Schaap and Bouten, 1996; Schaap et al., 1998; Tamari et al., 1996) – focusing dominantly on soil water retention; and,
2) Crop yield prediction - (Dai et al., 2011; Drummond et al., 1998; Drummond et al., 2003; Irmak et al., 2006; Khaki and Wang, 2019; Li et al., 2007; Liu et al., 2001; Niedbała, 2019; Pantazi et al., 2016; Park et al., 2005).

ANN are well suited to pedometrics for the development of PTF and SSPF, due to their strengths of modelling nonlinear relationships. Developments in PTF and SSPF have traditionally focused on use of linear modelling approaches (Wösten et al., 1995), where data availability has been limited. Advances in remote and proximal sensing however have increased the spatial resolution of environmental data used in the prediction of soil properties, and therefore, nonlinear ML approaches such as ANN should become advantageous. This has been recognised in the literature, where investigation of ANN approaches for PTF development have yielded improved performance over linear methods (Amini et al., 2005; Besalatpour et al., 2012; Koekkoek and Booltink, 1999; Sarani et al., 2016; Schaap et al., 1998).

The extensive use of ANNs for site-specific crop yield prediction is a key area of research which ANNs are inherently suited to, due to the highly complex and nonlinear relationships that exist between soil properties, climatic variables and yield. Understanding the key relationships between soil properties and crop performance is fundamentally useful as this provides a basis to inform site-specific management for system optimisation. Previous investigations have achieved acceptable yield predictive success during validation based upon the residual mean square error (RMSE) and coefficient of determination (R^2): RMSE=20% (Liu et al., 2001), RMSE=14.2% (Irmak et al., 2006), R^2=0.82 (Pantazi et al., 2016), R^2=0.84 (Dai et al., 2011). Whilst attempts have been made to apply the network in identifying sensitive

variables to infer the required system change to increase yield (Dai et al., 2011; Liu et al., 2001), these are largely constrained to surface conditions, with no attempt being made to incorporate depth-specific information. They therefore provide limited capacity to identify the influence of subsurface constraints, which largely limit crop production (Orton et al., 2018; Rengasamy, 2002), to subsequently estimate a yield response due to amelioration. Often soil depth data was not available for such investigations, due to the difficulties and costs associated with acquisition (Dai et al., 2011; Irmak et al., 2006; Liu et al., 2001; Pantazi et al., 2016).

Table 2.1. Summary of work pertaining to the application of ANN techniques for crop prediction.

Author	*ML Methods*	*No. of predictor variables*	*Predictor Variables*	*Dataset Size*	*Model Performance*	*Comments*
(Liu et al., 2001)	ANN	15	Soil ph, applied N fertiliser, soil P, soil K, soil organic matter, growing degree days, genetic potential, may rainfall, June rainfall, early July rainfall, late July rainfall, august rainfall, antecedent rainfall, planting density, rotation factor	360	RMSE = 20%	A a feed-forward back-propagation ANN to approximate the nonlinear relationship between corn yield and soil, management and climate variables. Model trained and tested using data obtained from a single scientific plot site. A sensitivity analysis identified the model parameters that resulted in maximised yield, discovering late yield was most sensitive to late July rainfall. Yield optimisation was not undertaken to globally optimise all model parameters.
(Drummond et al., 1998)	ANN	7	Soil ph, organic matter, P, K, CEC, topsoil depth, elevation	344	Training SE = 135kg.ha Testing SE = 277kg/ha	A feed-forward ANN to predict Soybean yield variability for a single year on a 36 ha site, exclusively using soil parameters as predictor variables. Multiple training algorithms, identifying resilient-propagation as superior. Predictive performance of the developed model was significantly reduced at the limits of the target data, with the network overestimated low yielding observations, and underestimated high observations. The cause was postulated to be a lack of data pre-treatment.
(Irmak et al., 2006)	ANN	14	Elevation, slope, wetness index, soil P, soil K, soil pH, Soil CEC, May rainfall, June rainfall, Early July rainfall, Late July rainfall, August rainfall, Soybean cyst nematode population, Weed density	120-543	Training: up to R^2=0.87 Testing: up to R^2=0.68	A feed-forward, back-propagation ANN to predict the spatial variability of soybean yields across different locations and years. Three cases were investigated, namely: i) predicting within-field variability in independent years, ii) predicting yield variability at independent sites, and iii) identifying crop stress factors that attribute to yield variability. Results showed that ANNs developed to predict yield variability should be field-specific, as model errors increased when making predictions outside of the geographical bounds of the training dataset, at independent sites.
(Dai et al., 2011)	ANN and MLR	10	Soil moisture and EC at sowing, seedling, squaring, flowering and maturity growth stages	108	ANN Training R^2=0.85 ANN Testing R^2=0.84 MLR training R^2=na MLR testing = R^2=0.70	Comparison of an ANN and MLR approach to predict sunflower yield based on soil moisture and soil salinity at different times during the growing season. ANNs had a higher predictive performance over the MLR.
(Niedbała, 2019)	ANN	21	Total precipitation and average air temperature for 5 time periods during growing season, N fertilizer (2 applications), P_2O_5 , K_2O, MgO, SO_3, B, Cu, Mn, Mo, Zn fertiliser applications	291	Training: R^2=0.92 Testing: R^2=0.91	Feed-forward back-propagation ANN trained for early prediction of winter rapeseed before harvest. The training dataset consisted of average crop yields from 291 sites over a period of 10 years. The developed model did not incorporate directly measured soil properties, but instead climate and management data (fertiliser applications). A sensitivity analysis indicated that crop yield was most sensitive to air temperature between January and April of the growing season.

Table 2.1. continued…

Author	*ML Methods*	*No. of predictor variables*	*Predictor Variables*	*Dataset Size*	*Model Performance*	*Comments*
(Khaki and Wang, 2019)	Deep ANN, Shallow ANN, Lasso, RT	707	627 genetic markers of plant genotype, precipitation, solar radiation, snow water equivalent, maximum temperature, minimum temperature and vapour pressure for each month, clay, silt, sand, available water capacity, soil pH, organic matter, CEC, soil saturated hydraulic conductivity	148,452	Training RMSE = 11.64 Testing RMSE = 13.94	Various ML techniques trained to predict the yield of various corn varieties for an independent season (2017). Training involved various corn hybrids grown at a number of geographic locations between 2008 and 2016. A deep ANN with 21 hidden layers produced superior results over shallow ANN, least absolute shrinkage and selection operator (Lasso) and regression trees (RT).
(Park et al., 2005)	ANN, GLM, RT	22	Soil pH, Organic matter, N-total, Soil P, Soil K, Soil Na, Soil Ca, Soil Mg, clay, silt, sand and 11 separate soil ameliorant treatments	720	Pearson's r	Various ML techniques trained to predict corn yield at various locations that are geographically spread. Training soil data was obtained from 720 agronomic trials with 11 applied treatments each. Optimal model performance was obtained by a regression tree model, with further work required to refine the ANN structure to improve predictions.
(Pantazi et al., 2016)	Counter-propagation ANN (CP-ANN), XY-fused networks (XY-Fs), SKNs	9	NDVI, Soil Ca, Soil CEC, Soil MC, soil organic carbon, soil P, soil pH, soil N (predicted using a partial least squares regression based on on-the-go vis-NIR spectroscopy and 60 direct measurements.	8798 (kriged data)	Testing ANN = R^2=0..783 Testing SKN = R^2=0.817 Testing X-Y-Fs = R^2=0..809	Prediction of crop yield variability for a single site and season, using various soil properties predicted from proximally sensed vis-NIR data. A total of 60 soil cores were used to calibrate the vis-NIR information to predict the soil properties using a partial least squares regression (PLSR) soil spatial prediction function (SSPF). The models were developed to predict discrete soil classes, as opposed to continuous values. The Supervised Kohonen Network (SKN) based on cross-validation training was most optimal for spatial yield prediction.

Table 2.2. Summary of work pertaining to the application of ANN techniques for soil water retention prediction

Author	*Predictor variable*	*Dataset size*	*Performance*	*Comments*
(Baker and Ellison, 2008)	clay (%), silt (%), sand (%), BD, OM (%)	2764	RMSE = 0.05-0.08 (m^3/m^-)	Ensemble method for ANN development, whereby multiple ANNs are integrated during predictions. Ensemble method required less data for training compared to traditional single ANN
(Koekkoek and Booltink, 1999)	BD, OM (%), clay (%), silt (%), sand (%) + matric potential and upper and lower boundary of pedology class	343	RMSE 0.026-0.0476 m^3/m^{-3}	Three ANNs each with different predictor variables to observe the effects of feature selection on model performance. Improved performance of the ANNs over previous regression approaches, however the magnitude was not significant
(Merdun et al., 2006)	clay (%), silt (%), sand (%), BD, OM (%)	195	R2 0.44-0.952	ANN achieved better predictions over multiple-linear regression, when only comparing the R^2, however this difference was not significant ($p > 0.05$)
(Minasny et al., 2004)	clay (%), silt (%), sand (%), BD, OM (%), saturated water content	310	RMSE = 0.69-0.81 m^3/m^{-3}	RMSE of water content and unsaturated hydraulic conductivity predictions of ANN were superior over the published *Rosetta PTF*. Most sensitive to sand content and saturated water content. The majority of soils were of a low clay content. The model had not been validated on an independent dataset.
(Minasny and McBratney, 2002)	clay (%), silt (%), sand (%), BD)	862	RMSE = 0.04 m^3/m^{-3}	A *neuro-m* method, whereby the parameters of the soil hydraulic model were predicted to optimise the PTF to match the observed and measured data.
(Minasny et al., 1999)	clay (%), silt (%), sand (%), BD, saturated water content, mean particle-size	842	RMSE = 0.57 m^3/m^{-3}	Improved validation using an extended nonlinear regression approach over an ANN, despite training errors being comparable. Concluded that more interpretable models such as ENR are often preferable over black-box approaches such as ANN due to better interpretability
(Schaap and Bouten, 1996)	clay (%), silt (%), sand (%), BD, OM (%)	204	RMSR 0.02 m^3/m^{-3}	Particle size distribution was most influential on the shape of the water retention curve whilst BD and organic matter were less important.
(Schaap et al., 1998)	clay (%), silt (%), sand (%), BD, OM (%), porosity, gravel content and soil horizons	620	RMSE 0.06 m^3/m^{-3}	Differenent sets of training features investigated. The larger number of features provided the best result. Better performance achieved using the ANN over published pedotransfer functions for soil water retention.

Nonlinear ML methods, such as ANN, do not perform well when presented new data that is outside the bounds of training data. This is due to the absence of structural assumptions that do not continue past the bounds of training. Irmak et al. (2006) observed this limitation when attempting to make yield predictions at a site that was independent of the sites used in training. The poor predictive performance suggested that the developed models were site-specific and could not be applied to make predictions at spatially different locations, where the covariates may be outside of the bounds of that in the training dataset. Exploring site-specific soil-crop interactions requires a local calibration of the model to best reflect the inherent site characteristics, or substantial datasets over greater geospatial variability such that the global structure can be determined. The former is a viable technique for using a ANN within modern day agriculture. However, the data density required to achieve this local calibration must be investigated, as there is a paucity of this information in the reviewed literature.

The ability for ANN to provide improved performance over linear methods is inherently dependent on the complexity of the modelling problem. Whilst improved predictive performance can be achieved using ANN (Amini et al., 2005; Besalatpour et al., 2012; Qiao et al., 2010a; Sarani et al., 2016) , there exists situations where linear methods may provide superior results (Park et al., 2005). It is also prudent to test ANN models against other methods to ensure the most appropriate modelling approach for the problem is applied.

A gradient-descent based backpropagation method is the most widely used training algorithm for ANN development (Besalatpour et al., 2012; Dai et al., 2011; Drummond et al., 2003; Hopmans et al., 2003; Irmak et al., 2006; Liu et al., 2001; Qiao et al., 2010a). Whilst gradient descent provides fast convergence, it is prone to finding local optimal solutions during training. The final solution is often dependent on the location within the feature space from which the network is initialised (Hopmans et al., 2003). Convergence to locally optimal solutions is common when the number of features is large in comparison to the number of training observations, or when the feature space is highly complex. Convergence to local optima can be assessed by iterating the model to observe changes in training and validation performance (Iyer and Rhinehart, 1999; Park et al., 1996). Pre-processing methods such as feature scaling or dimensionality reduction can be employed to reduce the likelihood of local optimum convergence (Baldi and Hornik, 1989; Wessels and Barnard, 1992).

A distinct limitation of ANNs is the requirement for complete datasets when providing inference (Ennett et al., 2001). This presents a large problem for agricultural modelling

purposes, as missing or incomplete datasets are extremely common, particularly when attempting in incorporate management information. Capturing this information to provide complete datasets therefore becomes a highly arduous task, and realistically, may not be achieved in the short-term. Other approaches to represent this missing management data via uncertainty metrics may provide a better option in the interim. Furthermore, ANNs do not attempt to model the mechanistic relationships between input and output variables, which may provide limitations for some modelling applications, especially where trends are counterintuitive to the domain knowledge relationship between a single variable and the output. An example of this is the fact that it is quite well established that subsoil constraints are less important in years with high in-crop rainfall, as compared to the inverse, which is a function of the ability to satisfy the crop from the topsoil and remove the subsoil impact for that data instance (Orton et al., 2018). Where sufficient training data exists, these mechanistic relationships may be explored using the model without reliance on domain knowledge. However, in situations of limited data availability (such as that for modelling temporal variability in soil constraint interactions), a mechantistic understanding of the relationships is required, but cannot be interrogated within the ANN, therefore meaning that domain knowledge is important to the interpretation of ANN output, particularly in data limiting environments.

Whilst ANNs provide some advantages over linear methods, they have key shortfalls and are limited by: overfitting in data limiting environments, susceptibility to local optimum convergence, inability to capture qualitative information, and inability to interrogate mechanistic relationships. Therefore, ANNs are not suited to all modelling problems presented in agriculture, and their application should be carefully structured to ensure appropriate fit to the given problem. In the context of soil constraint management, ANN have displayed opportunity to explore sensitive variables and infer the changes required to optimise yield (Liu et al., 2001), however this has been largely constrained to surface conditions, with no attempt being made to incorporate depth-specific information. Where data is sufficient, opportunity exists to apply ANNs with soil-depth information to better explore subsurface constraints, which are known to be large limiting factors of crop production (Orton et al., 2018; Rengasamy, 2002).

2.4.2. Support vector machines

2.4.2.A. Theory of support vector machines

Derived from statistical learning theory of Vapnik (1995), SVMs are considered a modern supervised ML technique which has been extensively used for image recognition, due to their accuracy in classification problems. Whilst they are commonly recognised as a classification and clustering tool (i.e. separating data points into discrete groups), their use has been extended to that of regression (i.e. prediction of continuous values based on observations within the data) (Smola and Schölkopf, 2004). A key strength of SVMs is that they employ Structural Risk Minimisation (SRM) which not only ensures a global optimum in the solution (Cristianini and Shawe-Taylor, 2000), but also prevents overfitting, as the focus is on minimising a bound on a risk function, as opposed to minimising training error (ANN approach) (Karimi et al., 2008). Further to that, they require less learning parameters to be defined by the modeler. SVMs for regression are formulated as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \langle w. x_i \rangle + b \qquad \text{Equation 2.5.}$$

where $<w. x_i>$ denotes the dot product of w and x_i, x_i is the feature space and b is the scalar bias. By adopting the ε-intrinsic loss function (Vapnik, 1995) and introducing slack variables (ζ_i, ζ_i^*) to quantify prediction errors, a constrained optimisation problem is formed, in which ε deviation must be maximised whilst the regularised loss minimised (refer to Figure 2.2.). The problem becomes:

$$\text{minimise} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*), \qquad \text{Equation 2.6.}$$

$$\text{subject to} \begin{Bmatrix} y_i - \langle w, x_i \rangle - b) \leq \zeta_i + \zeta_i^* \\ \langle w, x_i \rangle + b - y_i \leq \zeta_i + \zeta_i^* \\ \zeta_i + \zeta_i^* \geq 0, i = 1, \ldots, n \end{Bmatrix}$$

where C determines the trade-off between the flatness of f (model complexity) and the degree to which deviations larger that ε are tolerated.

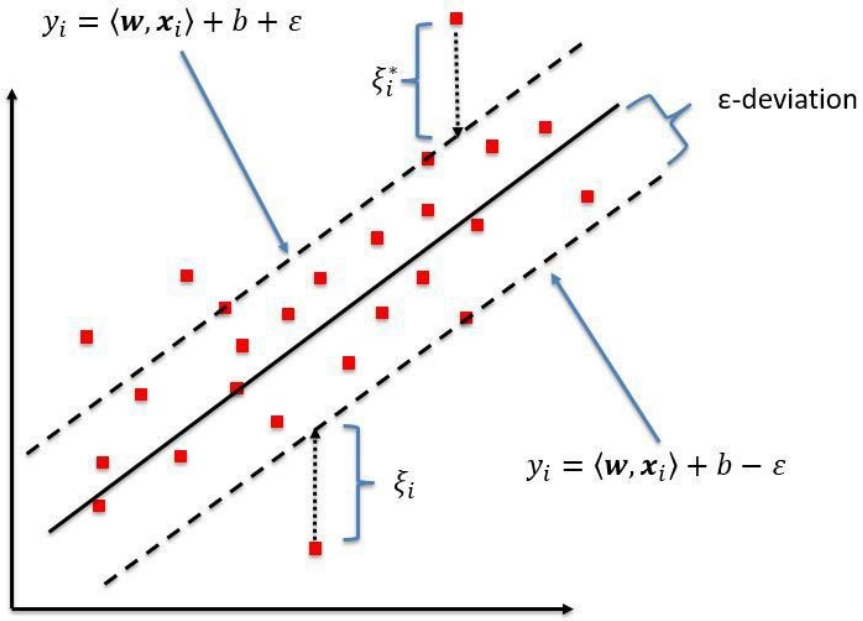


Figure 2.2. Fitting of the linear hyperplane for SVM regression using a 2-dimensional example. Support vectors, y_i are fitted to maximise the ε-deviation whilst simultaneously minimize regularized loss. Source: Kleynhans et al. (2017).

After introducing two Langrangian multipliers (α_i, a_i^*) and a Kernel function ($K(\boldsymbol{x}_i, \boldsymbol{x})$) to solve the convex quadratic problem of Equation 2.5. can be modified to give formulation of the final regression function (Equation 2.7.).

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (\alpha_i - a_i^*).K(\boldsymbol{x}_i, \boldsymbol{x}) + b \qquad \text{Equation 2.7.}$$

The primary purpose of the kernel (also known as the kernel trick) is to transform the input vector into a high-dimensional space such that the support vectors can be located to represent the data. Whilst many kernel functions exist (e.g. polynomial, sigmoid, guassian), the radial basis function (RBF) is the most widely used, as it is simple to use, requires just one hyperparameter (Li et al., 2008) and has fewer numerical difficulties, as it can only range from 0-1 (Hsu et al., 2003). However, RBF is not always suitable when there are many features (Hsu et al., 2003). The RBF is given as that presented in Equation 2.8:

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = \exp\left(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}\|^2\right) \qquad \text{Equation 2.8.}$$

Where γ is a scalar kernel parameter. In such a model, three parameters are therefore required for training, namely ε, C and γ which are selected via a parameter search (Hsu et al., 2003). The suggested ranges of this parameter search are ε (0-0.2), C (1-1x10^8) and γ (0.01 – 2) (Üstün et al., 2005). The optimum parameter values should result in the lowest RMSE of the model.

2.4.2.B. Adoption of support vector machines in agriculture

Application of SVMs for agricultural modelling purposes have been focused on two distinct areas, namely:

1) Pedotransfer development (Jafarzadeh et al., 2016; Khlosi et al., 2016; Lamorski et al., 2008; Twarakavi et al., 2009; Were et al., 2015)
2) Prediction and classification using remotely sensed data (Camps-Valls et al., 2003; Gill et al., 2006; Gualtieri and Cromp, 1999; Rumpf et al., 2010)

The application of these methods is summarised in Table 2.3.

SVMs offer an alternative modelling approach to PTF development, as compared to traditional methods, such as multiple linear regression, ordinary least squares and partial least squares. This is due to their ability to detect complex nonlinear patterns that are known to exist between soil variables (McBratney et al., 2003). PTFs are often developed on small datasets, where gradient-based methods, such as ANN, become at risk to convergence at a local optimum solution. The ability for SVMs to guarantee a global solution during training (Cristianini and Shawe-Taylor, 2000; Vapnik, 1995), makes them an attractive modelling tool for PTF development in data limiting environments. This is achieved due to the structural risk minimisation (SRM) approach SVMs employ during training (Vapnik, 1995). Whilst avoiding convergence to a local optimal solution may improve model performance and repeatability, it does not guarantee the model is appropriately fitted to the data, nor does it guarantee superior performance over other methods. Therefore, SVMs are still susceptible to overfitting, which should be considered for any modelling problem.

SVMs are well suited to providing inference out of high-dimensional datasets, such as that presented by remotely sensed hyperspectral technologies. This has led to their application for classification and predictive purposes in this space (Camps-Valls et al., 2003; Gualtieri and Cromp, 1999; Rumpf et al., 2010). The application of the kernel transformation into a higher dimensional space for locating the separating hyperplane ensures SVMs do not suffer from the

Hughes effect[1] (Houghes, 1968) unlike more traditional methods such as linear regression. Whilst this allows for all features to be used in training (Gualtieri and Cromp, 1999), it still may be advantageous to undertake data pre-processing, such as principal component analysis (PCA) to remove noise within the data (Chan et al., 2007; Rumpf et al., 2010).

SVMs offer similar advantages to ANNs, due to their ability to detect nonlinear patterns in high-dimensional data environments. This has resulted in the two methods often being directly compared in the literature for a given modelling problem (Gill et al., 2006; Jafarzadeh et al., 2016; Lamorski et al., 2008; Twarakavi et al., 2009; Were et al., 2015). In general, greater predictive accuracies have been achieved using SVM methods (Gill et al., 2006; Jafarzadeh et al., 2016; Lamorski et al., 2008), however, the magnitude of increased predictive performance does not warrant a generalised selection of SVMs as a superior modelling method for these applications. Instead, it would be advantageous to apply both SVM and ANN methods to the same modelling problem in an exploratory manner to identify the superior method for the specific modelling problem.

Whilst SVMs are a powerful nonlinear ML method, offering the advantage of avoiding local optimums, they can be computationally exhaustive in large datasets, (Eitrich and Lang, 2006; Momma and Bennett, 2002; Suykens and Vandewalle, 2000). This means that convergence time during training can be large, especially in the case of high-dimensional data (LeCun et al., 1995). Therefore, they may not be appropriate for situations where fast computations are required (e.g. on-the-go weed detection). The ability of SVMs to process high-dimensional data is a key advantage, however, similar to ANNs, they can be susceptible to overfitting (Han and Jiang, 2014), which should remain a key consideration when determining the quality of predictions. Improvements in model fit have been achieved by appropriately selecting the kernel function and parameters (Chuang et al., 2002).

[1] Phenomena of increased predictive performance to an optimal number of features, whereby performance is reduced thereafter by adding more features to describe the data. Referred to also as *The curse of dimensionality.*

Table 2.3. Summary of work pertaining to the application of ANN techniques in agriculture.

Author	*Prediction type*	*Kernel*	*Accuracy*	*Comments*
(Kovačević et al., 2010)	Classification and regression	Guassian	63% for classification. R^2 up to 0.94 for regression	Classification SVM (classifying soil types) and regression SVM (PTF). 9 soil types classified across a 1271 ha 63% accuracy. PTF SVM to predict t10 soil properties, with accuracy obtained for pH with an R^2=0.94.SVM models were compared against OLS and PLS models, with greater prediction accuracy being achieved for SVM.
(Rumpf et al., 2010)	Classification	RBF	Up to 97%	SVM model for early detection of plant disease based on remotely sensed hyperspectral information (25 spectral bands). The SVM was compared against an ANN model, with the SVM displaying slightly improved classification results. PCA was shown to greatly improve classification.
(Lamorski et al., 2008)	Regression	RBF	R^2 = 0.66–0.99 (depth dependent)	a PTF SVM model to predict soil hydraulic conductivity at multiple depths for the development of soil water retention curves. The SVM model was compared against a similar ANN model, with slightly improved predictions being obtained from the SVM model.
(Were et al., 2015)	Regression	SMO	R^2 = 0.64	PTF to predict soil organic carbon across a 650 km^2 region using 11 chemical and physical properties as predictor variables. SVM compared against an ANN, with slightly improved results being achieved from the SVM. The improved predictive performance of the SVM could not definitively identify SVMs are the superior approach.
(Kaundal et al., 2006)	Regression	RBF	R^2 = 0.61 (average)	SVM developed to predict the severity of crop disease using weather data obtained during key growth stages.
(Khlosi et al., 2016)	Regression	RBF	R^2 up to 0.75	SVM PTF to predict soil hydraulic conductivity at multiple depths to develop water retention curves. The SVM was compared with an ANN and MLR model, with the SVM model displaying improved results. The model was trained using only 54 samples, therefore overfitting was likely.
(Twarakavi et al., 2009)	Regression	RBF	RMSE down to 0.05	SVM PTF to predict soil hydraulic conductivity at multiple depths in the development of soil water retention curves using directly measured soil physical properties. SVM displayed improved results over a similar ANN, however, this was thought to be explained by the multi-model approach of the SVM, rather than the model itself.
(Jafarzadeh et al., 2016)	Regression	RBF	RMSE 2.796 Cmol kg^{-1}	SVM PTF for predicting soil CEC based on clay, silt, sand, gypsum and organic matter content. The SVM was compared against a similar ANN, with the ANN model displaying slight improvements. The magnitude of this difference was however not large enough to definitively select a superior model. Both models predicted poorly towards the bound of the training data.
(Gualtieri and Cromp, 1999)	Classification	Quadratic	Up to 96% classification	Classification SVM to detect crop types in individual fields using hyperspectral data with 128 bands. Classification accuracy of 87% for 16 classes, and 96% for 4 classes was chieved.
(Camps-Valls et al., 2003)	Classification	RBF and polynomial	Up to 98% classification	SVM classification model to classify crop types based on hyperspectral data with 128 bands. A polynomial kernel was able to achieve best modelling performance, however no indication toward overfitting was investigated for the small training dataset.
(Gill et al., 2006)	Regression	RBF	R^2 = 0.67	SVM model to future predict soil moisture conditions at 4 and 7 days in advance. Observed soil moisture conditions at time t and weather parameters between t and t+4&7 were used as predictor variables. The model was directly compared against a similar ANN, with the results obtained from the SVM being superior.

2.4.3. Clustering

2.4.3.A. Theory of clustering techniques

Clustering is an unsupervised ML technique that partitions ostensibly unstructured datasets into discrete groups (clusters) based on feature similarity (Rodriguez and Laio, 2014). Clustering data into discrete groups facilitates simpler and more effective management decisions, as the variability within the management unit (cluster) is reduced. Whilst a multitude of clustering algorithms exist, they can be broadly grouped into four main categories based on their applied algorithm (Fahad et al., 2014):

1) Partitioning-based,
2) Hierarchical-based,
3) Density-based, and
4) Model-based.

These four types of algorithms employ different optimization strategies, and their ability to accurately cluster data is dependent on the underlying structure of that data (see Figure 2.3). A brief overview of these four clustering algorithms is given below for each of these, with reference to Figure 2.3. For the purpose of explanation, *n* represents the number of observations within the dataset, and *k* is the number of clusters.

Partitioning-based approaches derive data similarity by assessing the distance between each data point and the centroid of its corresponding cluster. In the case of k-means clustering (most simple and widely used partitioning based algorithm), the aim is to locate the centroids of each cluster such that the intra-cluster variance is minimised using the squared error function (Equation 2.9.):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \qquad \text{Equation 2.9.}$$

K-means clustering involves the following steps:

1) Residence cluster of *k* centroids within the data feature space
2) Assign each data point to the closest centroid by Euclidean distance

3) Calculate the new centroids of each cluster by computing the mean location of each data point and re-assign each centroid to new location
4) Repeat steps 2&3 until convergence

Partitioning-based methods require the user to define the number of clusters, k, prior to computations, and are highly sensitive to outliers, as such data must be assigned to a cluster without consideration of its appropriateness. Random initialisation of the centroid locations reduces the model's repeatability in finding a global solution (Kassambara, 2017). Whilst partitioning-based algorithms are often computationally very efficient, their search to find clusters with similar Euclidean distances between each point and its corresponding centroid may result in clusters being ill-defined (Figure 2.3). This is particularly common where the data structure between clusters is not consistent.

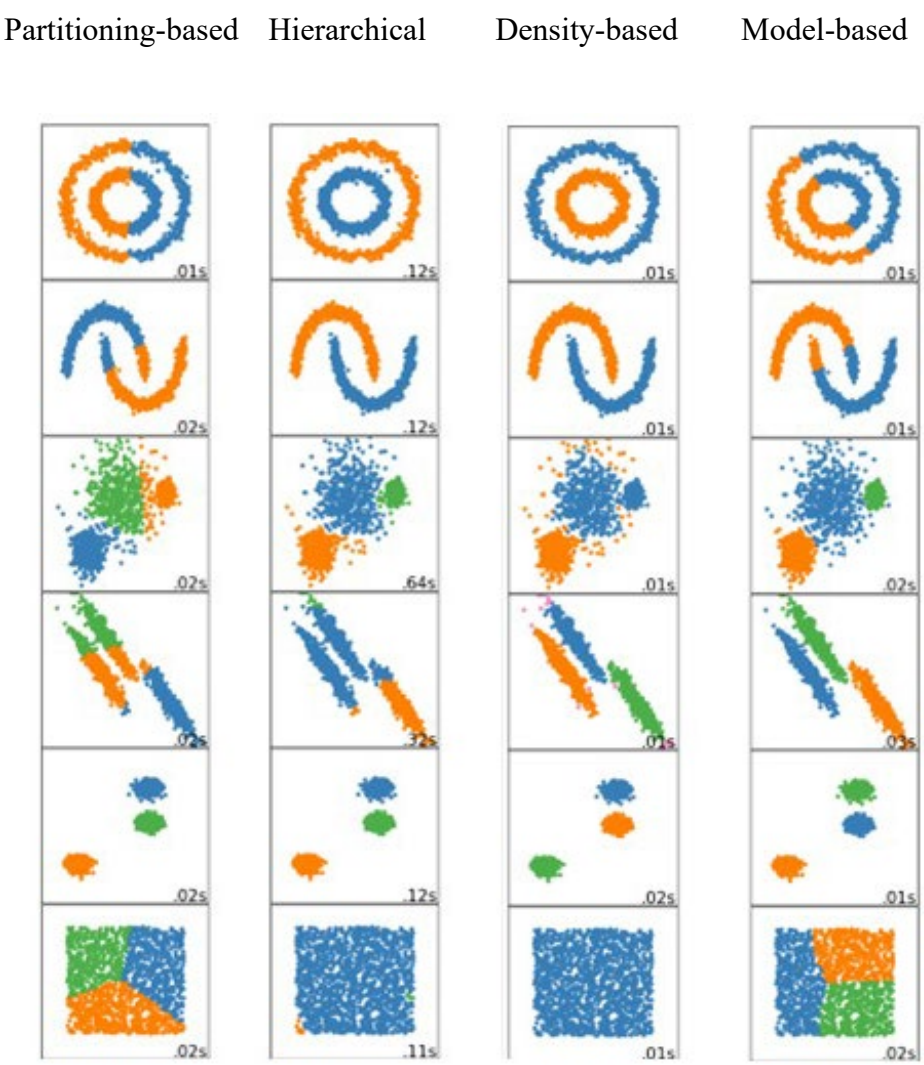


Figure 2.3. Examples of 2-dimensional clustering of 6 different datasets using partitioning-based, hierarchical, density and model-based approaches. Computational time for each method to cluster each dataset is presented in the lower-right corner of each illustration in units of seconds. Source: Seif (2018).

Hierarchical-based approaches iteratively separate data into clusters that have a predetermined ordering from top to bottom and can be represented as a dendrogram. Hierarchical clustering can either be agglomerative (i.e. initially $n=k$, with each iteration combining two clusters based on *similarity* until one cluster remains) or divisive (i.e. initially all observations are assigned to one cluster, with each iteration separating a cluster based on dissimilarity until $n=k$), as depicted

in Figure 2.4. Visual representation of the cluster process for agglomerative (left) and divisive (right) hierarchical clustering techniques. Modified from Stephanie (2016)

For agglomerate methods, data/cluster similarity is determined by the Euclidean distance between each instance in the feature space, either by single linkage (i.e. the distance between the two closest observations in each cluster), complete linkage (the distance between the two furthest observations in each cluster), or average linkage (the average distance between all points in each cluster). Divisive clustering recursively selects the cluster with the largest variance, or sum of squared error (SSE) splits the data in such a way that the squared error of the split clusters is minimised. Hierarchical clustering is better equipped to identify clusters with varied data structures, as each point may be considered individually during the interactive pruning process.

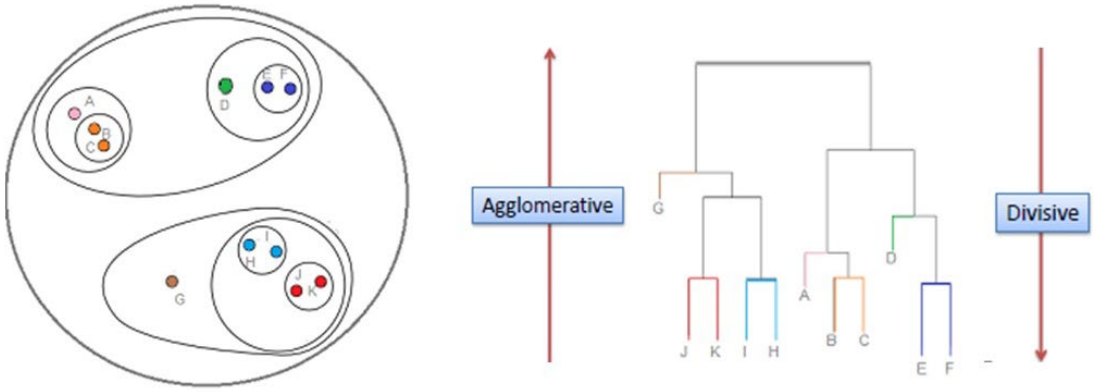


Figure 2.4. Visual representation of the cluster process for agglomerative (left) and divisive (right) hierarchical clustering techniques. Modified from Stephanie (2016)

Density based approaches, such as DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), DBCLASD (Xu et al., 1998) and DENCLUE (Hinneburg and Keim, 1998), aim to create clusters based on density (concentration of points) as opposed to Euclidean distance. The underlying assumptions are:

1) Clusters are dense regions in the data space, separated by regions of low object density; and,
2) A cluster is defined as a maximal set of density-connected points.

Density-based approaches do not require the user to specify the number of clusters in the dataset *a priori*, and are able to define clusters with varied data structures. They can also separate outliers (referred to as noise points) which do not fit into a cluster. This is achieved due to the methods' search which aims to cluster data within the bounds of a set 'density' that is specified by the user. Any data point that does not fit within the density requirement for a certain cluster, is assessed against its ability to fit within another cluster, or is considered an outlier.

Model-based approaches are based on the underlying assumption that the data was generated by some predefined mathematical model, the output of which can be described by probability distributions. Model-based methods are described as *soft clustering* techniques, as the boundary between each cluster is not definitive, with each data point represented by a probability of being assigned to each cluster. Such techniques iteratively locate the mean and standard deviation of each cluster's distribution, thus calculating the likelihood of each data point's assignment to each cluster. In the case of the commonly used model-based clustering technique — Gaussian-mixture model (Figure 2.5.) — expected maximisation (EM) is employed to optimise the model parameters μ_k (mean or centroid of cluster k), π_k (mixing coefficient for cluster k) and Σ_k (covariance for cluster k) such that model error is minimised. This involves two steps:

1. E Step: For each observation x_i, determine its assignment score to each Gaussian k:

$$\gamma(z_{ik}) = \frac{\pi_k P(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{m} \pi_j P(x_i|\mu_j, \Sigma_k)} \qquad \text{Equation 2.10.}$$

Where $\gamma(Z_{ik})$ is a measure of the 'responsibility', i.e. how much is Gaussian k responsible for the point x_i;

2. M Step: For each Gaussian k, update the parameters using $\gamma(Z_{ik})$.

$$\mu_k = \frac{1}{N_k} \sum \gamma(z_{ik}).x_i \qquad \text{Equation 2.11.}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{ik}) \qquad \text{Equation 2.12.}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{ik}).(x_i - \mu_k).(x_i - \mu_k)^T$$ Equation 2.13.

$$\pi_k = \frac{N_k}{N}$$ Equation 2.14.

where N is the total number of observations. This process is repeated until convergence to an optimal solution is achieved, where the change in error between each iteration is below a given threshold (see Figure 2.5). As the covariance is optimised for each cluster individually, clusters with varied data structures can be identified, as opposed to Euclidean based methods such as k-means clustering. This however is only limited to simple data structures, with the performance of model-based approaches declining as more advanced structures are presented.

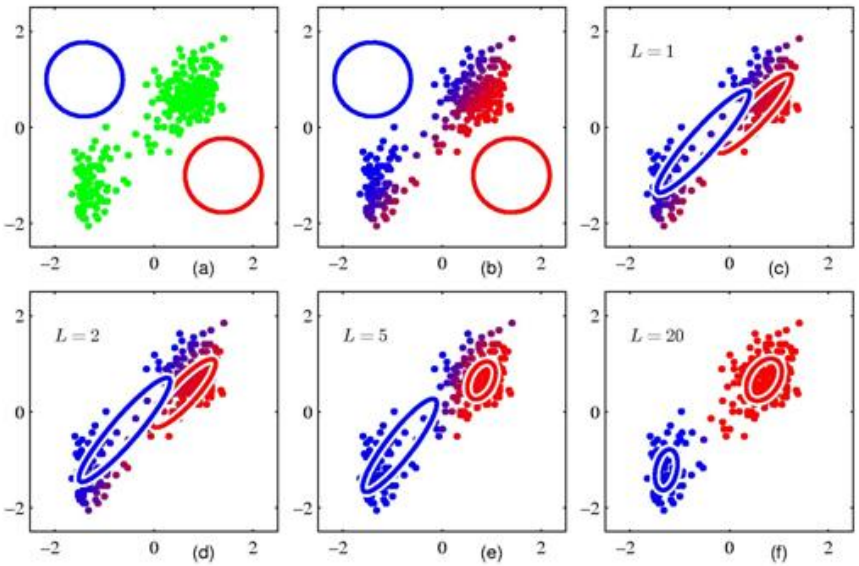


Figure 2.5. Illustration of the expected maximisation process to identify 2 Guassian model clusters for a 2-dimensional problem. Guassian models are randomly initialised (a) before converging on the identified clusters (b – f). Modified from Bishop (2006).

2.4.3.B. Adoption in agriculture

The adoption of clustering techniques in agriculture have been focused on problems pertaining to spatial data mining (Armstrong et al., 2007; Birant and Kut, 2007; Schoier and

Borruso, 2004), more specifically for the delineation of management zones for precision agriculture (Boydell and McBratney, 2002; Fleming et al., 2000; Fu et al., 2010; Li et al., 2007; Mehta et al., 2015; Ruß and Kruse, 2011), with further detail presented in Table 2.4. The concentration of work in this space is due to their suitability for handling spatial information in an attempt to identify discrete sub-regions within a landscape for management purposes. This occurs at a production level for variable rate soil recommendations (Fu et al., 2010), and at a government land management level for allocating funding (Khoshnevisan et al., 2015).

Spatial patterns in the landscape, both at the regional and sub-paddock scale, are known to be spatiotemporal dependent, due to the complex interaction between weather, soil and plant. Therefore, the time dimension must be considered when identifying management zones for precision agriculture. This has been recognised in the literature (Birant and Kut, 2007; Boydell and McBratney, 2002; Li et al., 2007), where the authors mapped spatial clusters, by inclusion of factors in the temporal domain. In an effort to map within-field management zones for precision agriculture, Boydell and McBratney (2002) identified that five years of yield data resulted in a Kappa Index of Agreement (KIA) of 75 % between years, however more than this number is preferred. Furthermore, this investigation was based on a single crop rotation (i.e.back-to-back-cotton), meaning a much greater number of years may be required for rotational farming systems to generate comparable temporal zones of similarity. This highlights the requirement of extensive temporal data in order to obtain robust spatial clusters.

Spatial datasets used in clustering can become quite large, due to advances in remote and proximal sensing which allows for spatially intensive data collection. This poses a computational challenge for clustering techniques. Parallel clustering provides the ability to increase computational efficiency for large datasets, as data can be broken down into discrete parts which can be solved concurrently, using separate computational resources (Ramesh et al., 2013; Zhao et al., 2013). Readers are referred to Ramesh et al. (2013) for practical methods to undertake parallel clustering. There has also been some evidence that the DBSCAN clustering algorithm has computational advantages in large spatial-temporal datasets (Birant and Kut, 2007; Ester et al., 1998; Zhou et al., 2000), due to a reduced number of iterations required during computations.

Table 2.4. Summary of work pertaining to the application of clustering techniques in agriculture.

Topic	*Algorithm*	*Author*
Clustering a soils dataset into discrete groups for statistical comparisons	Agglomerative hierarchical	(Gnatowski et al., 2010)
Delineation of site-specific management zones based on the spatiotemporal variability of soil EC	ISODATA (derivative of k-means)	(Yan et al., 2007)
Delineation of site-specific management zones based on measured soil properties	Agglomerative hierarchical	(Fleming et al., 2000)
Parallel algorithm to classify soil data sets	Parallel K-means	(Ramesh et al., 2013)
Clustering to identify groups of varying energy efficiencies to help inform practice change	C-means	(Khoshnevisan et al., 2015)
Identifying site-specific management zones based temporal satellite imagery	Fuzzy k-means	(Boydell and McBratney, 2002)
Nutrient management zone delineation in precision agriculture	Particle swarm optimization (PSO) – modified fuzzy k-means.	(Fu et al., 2010)
Delineation of site-specific management zones based on measured soil properties	Agglomerative hierarchical	(Ruß and Kruse, 2011)
Soil classification using fuzzy k-means clustering with extragrades	Fuzzy k-means with extragrades	(McBratney and de Gruijter, 1992b)
Estimating pedotransfer function prediction limits	Fuzzy k-means with extragrades	(Tranter et al., 2010)
Comparison of Clustering methods on a Precision Agriculture dataset	DBSCAN, OPTICS, Agglomerative, Divisive and COBWEB	(Mehta et al., 2015)
Characterising agricultural soil profiles	Expected maximisation and FarthestFirst	(Armstrong et al., 2007)
Clustering of spatiotemporal data	ST-DBSCAN (variant of DBSCAN)	(Birant and Kut, 2007)
Clustering in large spatial databases	MDBSCAN (variant of DBSCAN)	(Schoier and Borruso, 2004)

The k-means clustering technique is the most widely used in the literature (Boydell and McBratney, 2002; Fu et al., 2010; Khoshnevisan et al., 2015; McBratney and de Gruijter, 1992b; McBratney and Pringle, 1999; Ramesh et al., 2013; Ruß and Kruse, 2011; Tranter et al., 2010). The computational simplicity of k-means clustering contributes to its superior clustering speed for small-medium sized datasets (Mehta et al., 2015). Furthermore, the number of clusters, k, is the only parameter required for the model.

However, k-means clustering is limited by the adopted 'hard-clustering' approach which assumes discrete assignment of each observation to a single cluster. In reality, natural systems do not have discrete boundaries, and are instead continuous in nature. This is true for spatial variation of soil properties, which do not change abruptly at a given boundary, but instead continuously change across a landscape. The limitation of this assumption for predictive soil mapping has been recognised by De Gruijter and McBratney (1988) and McBratney and de Gruijter (1992b), who extended on the work of (Bezdek, 1975) to present a fuzzy k-means approach. This allows 'soft-clustering', whereby each observation is assigned a continuous membership value to every cluster, depending on the degree of similarity to other observations assigned to the respective clusters (McBratney and Odeh, 1997). The use of extragrades in fuzzy clustering within the context of agriculture was also introduced by McBratney and de Gruijter (1992b) to better identify outliers and separate them into their own cluster accordingly. The use of extragrades also reduces the sensitivity to outliers. This has since formed a large body of work surrounding the use of fuzzy clustering in pedometrics and soil classification (Bragato, 2004; Hanesch et al., 2001; McBratney et al., 2003; Odeh et al., 1992; Yang et al., 2011).

One of the major limitations of partitioning and model-based clustering approaches is the requirement for the number of clusters to be pre-specified (Pham et al., 2005). This may result in groups of data being forced into homogenous clusters, when in reality, they are inherently different. On the contrary, uniform data may also be separated to satisfy the defined number of clusters (Figure 2.3). Whilst methods to optimise k exist in the literature (Chiang and Mirkin, 2010), practical implementation of k-means clustering for zonal soil management commonly results in the clusters being selected subjectively, often influenced by the number of units the user is desiring to manage (e.g. selecting 4 agricultural management zones based on the number of samples a land

owner is willing to invest). Furthermore, partitioning based approaches have limited ability to define arbitrarily shaped clusters of varied densities and distributions within data, due to its basis of Euclidean distance. Whilst these approaches occupy a number of limitations, their ability to handle large, high-dimensional data in a fast and efficient manner (Budayan et al., 2009) lends them well to application of remotely sensed information within the context of agriculture.

Density-based and hierarchical approaches overcome the limitations of partitioning-based and hierarchical clustering by seeking to identify the optimum number of clusters based on the natural structure within the data. Furthermore, these approaches can separate outliers within the data, therefore identifying observations which may not be appropriately suited to the identified clusters. Whilst density-based approaches have previously been limited in identifying clusters of varied densities, recent advances provide opportunity to overcome this (Birant and Kut, 2007). However, hierarchical and density-based approaches are largely limited in handling high-dimensional data, as the feature space is often sparse (Jain, 2010; Milenova and Campos, 2002). Therefore, they may not be well-suited to spatial problems in agriculture, where high-dimensional data is present (e.g. hyperspectral remotely sensed data).

2.4.4. Bayesian belief networks

2.4.4.A. Theory of Bayesian belief networks

BBNs are often employed to analyse decision strategies under uncertain conditions (Aguilera et al., 2011; Varis, 1997) and in environments with incomplete data, making them highly advantageous for use in the context of agriculture (Pollino and Henderson, 2010). Due to their highly adaptive structure, BBNs have a wide range of applications, including automated monitoring, prediction, cause identification, classification and decision support (Drury et al., 2017). A series of agricultural applications is presented within Table 2.5. BBNs have many core strengths, which include, but are not limited to the following:

1) The ability to strengthen decisions by integrating qualitative and quantitative information from various sources when pure empirical data is lacking (Smith et al., 2007b);
2) The ability to update the model when new information becomes available (Henriksen and Barlebo, 2008); and,
3) They can run simulations with incomplete input data.

BBNs are often referred to as a probabilistic graphical models, as the models utilise probability to determine relationships and outcomes, and can be graphically represented. They consist of two components, namely a directed acyclic graph (DAG) and a set of conditional probability tables (CPTs). The DAG is represented by a series of nodes (variables) and directed arcs which represent the relationships between the nodes. The acyclic nature of the relationships means that information flow is unidirectional and therefore feedback loops cannot be represented, which can be a limitation in some instances (Drury et al., 2017). CPTs are used to quantitatively describe the strength of the conditional dependencies between the nodes within the DAG (Bashari et al., 2008) as discrete probabilities. This means that system variables cannot have continuous states and need to be in a discrete form, which can be considered as another limitation, depending on the data source and application. Subsequently, the temporal dynamics cannot be directly built into the model. An important consideration here is where some action (e.g. application of a soluble agronomic amendment) does not have an effect that occurs instantaneously, given some other condition (e.g. application to land of the amendment). This means that the model's assumption through the mathematics that 'action' instantaneously equates to 'effect' given 'condition' may not match the physical circumstance (e.g. dissolution of the amendment over time after application).

Information is passed through the network using Bayes' Theorem, represented as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{Equation 2.15.}$$

Where P(A) is the prior distribution of parameter A, P(B|A) is the probability of variable B, given existing knowledge of A and the posterior probability of A given B.

Three approaches exist for the initial development of a BBN, namely:

1) Relying solely on expert knowledge to construct both the physical network structure (identifying key nodes and directional information flow) and assign the conditional probabilities between nodes;
2) Both the network structure and conditional probabilities between nodes are machine learnt from data; or,
3) A hybrid approach where the network structure is constructed using expert knowledge and the conditional probabilities are machine learnt from available data (Tari, 1996)

A major benefit of learning within a BBN is that discrete nodes within the network which were originally trained using expert opinion can be retrained with empirical data if/when it becomes available, therefore building model robustness. The notion of using expert opinion in training is highly beneficial when no such data or poorly represented data exists on the functioning of a particular system variable/s. BBNs are therefore highly suited to agricultural situations, and as such, their use within this domain is reviewed below.

2.4.4.B. Adoption in agriculture

A key advantage of BBNs over empirical ML approaches is the ability to merge quantitative and qualitative data streams to provide inference within a system. This is reflected in the literature by a combination of expert opinion and data to provide inference (Table 2.5.). This is either achieved by using expert opinion to describe the conditional relationships within an empirical system, or by the incorporation of variables with qualitative states (Bi and Chen, 2011; Kristensen and Rasmussen, 2002; Smith et al., 2007b; Troldborg et al., 2013). This allows for the incorporation of qualitative variables which are difficult to physically measure (e.g. soil structure), but are known to be highly influential on system processes (e.g. water and nutrient cycling) and can be categorised.

The use of expert opinion for model learning allows for modelling in data limited environments, where empirical ML approaches struggle to reach convergence on a global solution. This is common in the literature (Farmani et al., 2009; Robertson and Wang, 2004; Smith et al., 2007b; Troldborg et al., 2013), where a panel of domain experts are utilised to identify dependency and independency between variables, as well as the direction of influence between these. The use of expert opinion is, however, constrained by the existing domain knowledge surrounding key system relationships, which is not available in some situations. Furthermore, expert opinion is only suited to small networks, as the number of conditional probability states for a given node exponentially increase with network size (e.g. a node with 4 input nodes, each with 4 states, requires 256 conditional probabilities for the CPT). Therefore, in the presence of high-dimensional data or where domain knowledge is lacking, empirical approaches to network learning are more suited (Chawla et al., 2016; Kristensen and Rasmussen, 2002; Robertson and Wang, 2004; Troldborg et al., 2013).

Table 2.5. Summary of work pertaining to BBN application in agriculture

Topic	*Structure Learning*	*CPT Learning*	*Application*	*Author*
Disease detection in individual dairy cows	Data based	Empirical data	Automated monitoring, prediction	(Steeneveld et al., 2010)
Crop disease detection	Expert Opinion	Expert opinion + literature	Automated monitoring, prediction	(Bi and Chen, 2011)
Insect outbreak detection	Expert Opinion	Expert Opinion	Automated monitoring	(Holt et al., 2006)
Predicting Crop energy yield	Expert Opinion	Data based + existing mathematical formula	Prediction	(Newlands and Townley-Smith, 2010)
Mapping areas of habitat suitability	Expert opinion	Expert opinion + literature	Classification	(Smith et al., 2007b)
Optimal management practice selection for weed presence	Expert opinion	Expert opinion, empirical data + mathematical formulae	Decision support	(Kristensen and Rasmussen, 2002)
Optimal management of groundwater contamination	Expert Opinion	Expert opinion + mathematical formulae	Decision support	(Farmani et al., 2009)
Selecting optimal irrigation practice	Expert Opinion	Expert opinion + literature + empirical data	Decision support	(Robertson and Wang, 2004)
Crop yield prediction	Expert Opinion	Empirical data	Decision support + Prediction	(Chawla et al., 2016) (Van Der Gaag et al., 2010)
Evaluating dispersive spoil rehabilitation approaches	Expert Opinion	Expert opinion + literature + empirical data	Decision support	Reardon-Smith et al. 2017
Soil compaction risk prediction	Expert Opinion	Expert opinion + literature + empirical data	Prediction and diagnosis	(Troldborg et al., 2013)

The ability of BBNs to account for uncertainty allows the precision of predictions to be represented. This is a key advantage over empirical ML approaches which express predictions as a mean value, thus potentially providing the user a false indication of model precision (Kristensen and Rasmussen, 2002). This allows the user to quantify the level of risk associated with management decisions based on model outcomes, which is a major limitation of current decision support tools (DST) in industry. Uncertainty can also inform the focus of future data collection by

identifying variables and relationships within the model that are not well understood (Troldborg et al., 2013).

BBNs reduce model parameterisation by utilising discrete states for variables. Traditional models often become complex, and although they might be powerful, they require a high degree of parameterisation that is not always possible, is inefficient, or exceeds logistical constraints. These variables are often omitted from the modelling exercise, despite their contextual relevance. Examples include representing crop sowing date as 'early', 'average' or 'late' when the exact date has not been recorded (Gu et al., 1994), or denoting soil type as 'cracking clay' or 'non-cracking clay' when an extensive description of soil attributes is not available and the attribute of 'cracking' has meaningful influence on the outcome variable (Smith et al., 2007). While it is possible to include complex structure into BBN models, this typically results in detailed CPTs which are difficult to populate well without training data. The simplest approach to this is to limit the number of discrete state values for each of the conditional and output variables, in lieu of obtaining more training data, meaning that the resolution of the model is reduced. Irrespective of this consideration, the discrete state value approach of BBN results in suitability for investigation of holistic and complex systems.

Agricultural decisions are often based on incomplete and missing information, which inhibits inference using empirical approaches. To simply wait for more information to become available (e.g. weather data) is not feasible, as the opportunity risk exceeds the risk of a poor decision as time continues. BBNs however can reason with partial information by relying on the statistical distributions of the trained nodes in the network, therefore allowing inference with incomplete data. As new data presents, 'belief updating' can be employed, whereby predictions and uncertainties are readjusted to represent the new information (Chawla et al., 2016). This allows for risk-based decision making, whereby the user can simultaneously compare prediction uncertainty with data availability in an effort to satisfy their risk appetite.

BBNs are highly applicable for modelling in the agricultural domain, largely due to their ability to integrate both qualitative and quantitative information, make inference with partial/missing information and provide a degree of confidence in the output. They are however limited in situations which are computationally complex (i.e. large number of variables with many possible states), where other ML approaches such as ANNs or SVMs may be better suited. Therefore, it is

likely that BBNs will be a useful tool for initial modelling in situations where empirical data is not yet available. As such data becomes available (e.g. due to increased sensor development/availability), there may be a transition from statistical to empirical models better suited to pure empirical data. Furthermore, there is opportunity for BBNs to be used as an integrated decision tool which may be able to make inference on information generated by other ML models.

2.5. Opportunities for constraint diagnosis and yield prediction

Each of the major sections has provided a table of information investigating the application of ML methods to agriculture for a variety of tasks. Considering these, and the mathematical reasoning behind the methods, there are numerous of research opportunities for constraint diagnosis and yield prediction within agriculture, considering the soil resource as the context:

ANNs and SVMs both have proven abilities to represent nonlinear relationships in a variety of agricultural situations (Behmann et al., 2015; Irmak et al., 2006; Lamorski et al., 2008; Minasny and McBratney, 2002). Whilst inherently different in their mathematical formulation and approach to regression, both techniques have achieved similar predictive accuracies when applied to the same modelling problem (Behmann et al., 2015; Jafarzadeh et al., 2016; Khlosi et al., 2016; Lamorski et al., 2008; Twarakavi et al., 2009; Were et al., 2015). We therefore see it imperative to investigate both methods simultaneously.

Of importance is the ability to apply these methods to model nonlinear soil-crop interactions, similar to that pursued by Drummond et al. (1998), Irmak et al. (2006) and Liu et al. (2001), however for the purpose of identifying site-specific yield limiting soil constraints. Opportunity exists to apply these models in simulating a yield response due to a soil amelioration strategy, therefore allowing for improved variable-rate (VR) recommendation advice with consideration towards the economics of application and likely return. This is a key missing link in the adoption of soil amelioration strategies in industry (Bennett and Cattle, 2014; Lobry de Bruyn and Andrews, 2016). Furthermore, these approaches should be extended to incorporate depth-specific soil information to separate surface constraints from subsurface constraints, which are known to be highly influential on crop performance (Dang et al., 2010; Orton et al., 2018). To our knowledge, this has not been attempted in the literature at an intensive spatial scale.

Whilst nonlinear approaches offer the ability to accurately identify complex soil-yield relationships, the lack of structural assumptions causes susceptibility to overfitting and poor generalisation (Briscoe and Feldman, 2011; Tu, 1996), particularly in data limiting situations. Therefore, in applying these techniques, direct consideration of these limitations is required to ensure their appropriateness over linear approaches for the given modelling problem. Methods to assess generalistion beyond the R^2 are further required, as this metric can be misleading (Alexander et al., 2015) and can provide false confidence of a model's performance. This is particularly concerning when model results inform decisions of large economic significance (i.e. soil amelioration). Furthermore, attention is required toward the interpretability of ML approaches to better assess model generalisation.

The cost of improved modelling complexity for nonlinear ML approaches is an increased training data requirement (Wani et al., 2019), which may not be satisfied in some agricultural modelling contexts. In these situations, generalisation can be improved by employing data augmentation techniques to artificially increase the sample size to reduce the risk of overfitting (Perez and Wang, 2017; Santoro et al., 2016). For spatial datasets, geostatistical and non-interpolation techniques such as OK (Burgess and Webster, 1980a) and SSPFe (McBratney et al., 2003) present the opportunity for improved generalisation of spatial predictions by increasing the number of spatial observations (Park et al., 2005).

Empirical ML approaches are limited in their ability to handle missing or incomplete data (Ennett et al., 2001) and incorporate qualitative relationships between system variables. They are therefore not well suited to integrating management information, which is known to be highly influential on soil function (e.g. compaction and random in-field traffic versus spatially controlled traffic`; (Bartimote et al., 2017; Bennett et al., 2017). Qualitative, probabilistic methods, such as BNNs offer opportunity within this context (Pollino and Henderson, 2010), due to their ability to integrate empirical data with expert opinion to make predictions with an expressed degree of uncertainty (Kristensen and Rasmussen, 2002; Troldborg et al., 2013). Whilst they are inherently useful for application with management information in uncertain situations, they are not well suited in the presence of high-dimensional information, and cannot predict continuous classes.

Each ML method investigated provides opportunity for improved spatial management and soil constraint diagnosis within agriculture. Whilst these are powerful for individual modelling

problems, consideration is required for the integration of these techniques to provide a holistic approach to data-driven decisions in precision agriculture. This hybrid approach will require the utilisation of empirical and probabilistic methods to better merge quantitative and qualitative data streams with the aim of providing useful insight to on-farm decision making.

2.6. Conclusion

Precision agriculture currently requires better understanding of the spatial variability of soil constraints at the sub-paddock level to achieve improved economic performance through targeted amelioration strategies. This will involve the use of various data sources and analytical techniques to not only map soil constraints, but provide inference toward constraint-yield interactions for improved management. Whilst this review has identified many data sources which can be used to develop soil constraint maps, it is not clear how much directly measured soil and data is required to build these models to provide useful soil insight. Perhaps more importantly, there is also no indication toward the cost-benefit relationship between collecting more soil data to obtain improved models and subsequent management decisions in relation to the cost of data acquisition. Furthermore. Furthermore, although ANNs, SVMs, clustering algorithms and BBNs display potential to model soil-yield interactions in an effort to identify the effects of soil constraints on crop yield, the best modelling approach and volume of data required to achieve model generalisation whilst avoiding overfitting appears to be specific to the modelling problem. These methods, and others, should be further investigated for their ability to identify sub-paddock soil-yield interactions with varying volumes of training data.

This review has also identified that qualitative information and missing data are common occurrences in agricultural modelling situations, and therefore empirical approaches may not always be appropriate to provide inference. Statistical approaches such as BBNs offer a solution to such problems, however their ability to provide insight for specific soil issues should be further investigated.

2.7. References

Adamchuk, V.I., Hummel, J., Morgan, M., Upadhyaya, S., 2004. On-the-go soil sensors for precision agriculture. Computers and electronics in agriculture 44(1), 71-91.

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environmental Modelling & Software 26(12), 1376-1388.

Alexander, D., Tropsha, A., Winkler, D.A., 2015. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. Journal of chemical information and modeling 55(7), 1316-1322.

Amini, M., Abbaspour, K.C., Khademi, H., Fathianpour, N., Afyuni, M., Schulin, R., 2005. Neural network models to predict cation exchange capacity in arid regions of Iran. European Journal of Soil Science 56(4), 551-559.

Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure, ACM Sigmod record. ACM, pp. 49-60.

Armstrong, L.J., Diepeveen, D., Maddern, R., 2007. The application of data mining techniques to characterize agricultural soil profiles, pp. 85-100.

Arslan, S., Colvin, T.S., 2002. Grain yield mapping: Yield sensing, yield reconstruction, and errors. Precision Agriculture 3(2), 135-154.

Atzberger, C., 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote sensing 5(2), 949-981.

Baharom, S.N.A., Shibusawa, S., Kodaira, M., Kanda, R., 2015. Multiple-depth mapping of soil properties using a visible and near infrared real-time soil sensor for a paddy field. Engineering in agriculture, environment and food 8(1), 13-17.

Baker, L., Ellison, D., 2008. Optimisation of pedotransfer functions using an artificial neural network ensemble method. Geoderma 144(1-2), 212-224.

Baldi, P., Hornik, K., 1989. Neural networks and principal component analysis: Learning from examples without local minima. Neural networks 2(1), 53-58.

Ballabio, C., 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151(3-4), 338-350.

Barnes, E., Clarke, T., Richards, S., Colaizzi, P., Haberland, J., Kostrzewski, M., Waller, P., Choi, C., Riley, E., Thompson, T., 2000. Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data, Proceedings of the Fifth International Conference on Precision Agriculture, Bloomington, MN, USA.

Bartimote, T., Quigley, R., Bennett, J.M., Hall, J., Brodrick, R., Tan, D.K., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: II. Economic and physiological analysis. Soil and Tillage Research 168, 133-142.

Bashari, H., Smith, C., Bosch, O., 2008. Developing decision support tools for rangeland management by combining state and transition models and Bayesian belief networks. Agricultural Systems 99(1), 23-34.

Behmann, J., Mahlein, A.-K., Rumpf, T., Römer, C., Plümer, L., 2015. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. Precision Agriculture 16(3), 239-260.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Journal of plant nutrition and soil science 168(1), 21-33.

Bennett, J.M., Cattle, S., 2014. Adoption of soil health improvement strategies by Australian farmers: II. Impediments and incentives. The Journal of Agricultural Education and Extension 20(1), 107-131.

Bennett, J.M., Roberton, S.D., Jensen, T.A., Antille, D.L., Hall, J., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: I. Heavy machinery impact on the soil resource. Soil and Tillage Research 168, 143-154.

Bentley, M.L., Mote, T.L., Thebpanya, P., 2002. Using Landsat to identify thunderstorm damage in agricultural regions. Bulletin of the American Meteorological Society 83(3), 363-376.

Besalatpour, A., Hajabbasi, M., Ayoubi, S., Afyuni, M., Jalalian, A., Schulin, R., 2012. Soil shear strength prediction using intelligent systems: artificial neural networks and an adaptive neuro-fuzzy inference system. Soil science and plant nutrition 58(2), 149-160.

Bezdek, J.C., 1975. Mathematical models for systematics and taxonomy, Proceedings of eighth international conference on numerical taxonomy, pp. 143-166.

Bi, C., Chen, G., 2011. Bayesian Networks Modeling for Crop Diseases. In: D. Li, Y. Liu, Y. Chen (Eds.), Computer and Computing Technologies in Agriculture IV: 4th IFIP TC 12 Conference, CCTA 2010, Nanchang, China, October 22-25, 2010, Selected Papers, Part I. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 312-320.

Birant, D., Kut, A., 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data & Knowledge Engineering 60(1), 208-221.

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Bishop, T., McBratney, A., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103(1-2), 149-160.

Boydell, B., McBratney, A., 2002. Identifying potential within-field management zones from cotton-yield estimates. Precision agriculture 3(1), 9-23.

Bragato, G., 2004. Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain. Geoderma 118(1-2), 1-16.

Briscoe, E., Feldman, J., 2011. Conceptual complexity and the bias/variance tradeoff. Cognition 118(1), 2-16.

Brus, D., Kempen, B., Heuvelink, G., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62(3), 394-407.

Budayan, C., Dikmen, I., Birgonul, M.T., 2009. Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. Expert Systems with Applications 36(9), 11772-11781.

Buhmann, M.D., 2003. Radial basis functions: theory and implementations, 12. Cambridge university press.

Burgess, T., Webster, R., 1980. Optimal interpolation and isarithmic mapping of soil properties: I. The semivariogram and punctual kriging. Journal of soil science 31(2), 315-331.

Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Soria-Olivas, E., Martín-Guerrero, J.D., Moreno, J., 2003. Support vector machines for crop classification using hyperspectral data, Iberian Conference on Pattern Recognition and Image Analysis. Springer, pp. 134-141.

Chan, A.B., Vasconcelos, N., Lanckriet, G.R., 2007. Direct convex relaxations of sparse SVM, Proceedings of the 24th international conference on Machine learning. ACM, pp. 145-153.

Chang, D.-H., Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural network. Remote Sensing of Environment 74(3), 534-544.

Chawla, V., Naik, H.S., Akintayo, A., Hayes, D., Schnable, P., Ganapathysubramanian, B., Sarkar, S., 2016. A Bayesian Network approach to County-Level Corn Yield Prediction using historical data and expert knowledge. arXiv preprint arXiv:1608.05127.

Chuang, C.-C., Su, S.-F., Jeng, J.-T., Hsiao, C.-C., 2002. Robust support vector regression networks for function approximation with outliers. IEEE Transactions on Neural Networks 13(6), 1322-1330.

Cressie, N., 1985. Fitting variogram models by weighted least squares. Journal of the International Association for Mathematical Geology 17(5), 563-586.

Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Dai, X., Huo, Z., Wang, H., 2011. Simulation for response of crop yield to soil moisture and salinity with artificial neural network. Field Crops Research 121(3), 441-449.

Dang, Y., Moody, P., 2016. Quantifying the costs of soil constraints to Australian agriculture: a case study of wheat in north-eastern Australia. Soil Research 54(6), 700-707.

Dang, Y.P., Dalal, R.C., Buck, S., Harms, B., Kelly, R., Hochman, Z., Schwenke, G.D., Biggs, A., Ferguson, N., Norrish, S., 2010. Diagnosis, extent, impacts, and management of subsoil constraints in the northern grains cropping region of Australia. Soil Research 48(2), 105-119.

DasGupta, B., Schnitger, G., 1993. The power of approximating: a comparison of activation functions, Advances in neural information processing systems, pp. 615-622.

De Gruijter, J., McBratney, A., 1988. A modified fuzzy k-means method for predictive classification.

Dlugoß, V., Fiener, P., Schneider, K., 2010. Layer-specific analysis and spatial prediction of soil organic carbon using terrain attributes and erosion modeling. Soil Science Society of America Journal 74(3), 922-935.

Drummond, S., Joshi, A., Sudduth, K.A., 1998. Application of neural networks: precision farming, Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on. IEEE, pp. 211-215.

Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J., Kitchen, N.R., 2003. Statistical and Neural Methods for Site–Specific Yield Prediction. Transactions of the ASAE 46(1), 5.

Drury, B., Valverde-Rebaza, J., Moura, M.-F., de Andrade Lopes, A., 2017. A survey of the applications of Bayesian networks in agriculture. Engineering Applications of Artificial Intelligence 65, 29-42.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.

Eitrich, T., Lang, B., 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. Journal of Computational and Applied Mathematics 196(2), 425-436.

Ennett, C.M., Frize, M., Walker, C.R., 2001. Influence of missing values on artificial neural network performance, Medinfo, pp. 449-453.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, Kdd, pp. 226-231.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1998. Clustering for mining in large spatial databases. KI 12(1), 18-24.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S., Bouras, A., 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing 2(3), 267-279.

Farmani, R., Henriksen, H.J., Savic, D., 2009. An evolutionary Bayesian belief network methodology for optimum management of groundwater contamination. Environmental Modelling & Software 24(3), 303-310.

Fiener, P., Auerswald, K., 2009. Spatial variability of rainfall on a sub ‐ kilometre scale. Earth Surface Processes and Landforms 34(6), 848-859.

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

Florinsky, I.V., Eilers, R.G., Manning, G., Fuller, L., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling & Software 17(3), 295-311.

Fu, Q., Wang, Z., Jiang, Q., 2010. Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO. Mathematical and computer modelling 51(11-12), 1299-1305.

Genc, H., Genc, L., Turhan, H., Smith, S., Nation, J., 2008. Vegetation indices as indicators of damage by the sunn pest (Hemiptera: Scutelleridae) to field grown wheat. African Journal of Biotechnology 7(2).

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines. JAWRA Journal of the American Water Resources Association 42(4), 1033-1046.

Gnatowski, T., Szatyłowicz, J., Brandyk, T., Kechavarzi, C., 2010. Hydraulic properties of fen peat soils in Poland. Geoderma 154(3-4), 188-195.

Gualtieri, J.A., Cromp, R.F., 1999. Support vector machines for hyperspectral remote sensing classification, 27th AIPR Workshop: Advances in Computer-Assisted Recognition. International Society for Optics and Photonics, pp. 221-233.

Han, H., Jiang, X., 2014. Overcome support vector machine diagnosis overfitting. Cancer informatics 13, CIN. S13875.

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

Hawkins, D.M., Basak, S.C., Mills, D., 2003. Assessing model fit by cross-validation. Journal of chemical information and computer sciences 43(2), 579-586.

Haykin, S., Network, N., 2004. A comprehensive foundation. Neural networks 2(2004), 41.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124(3), 383-398.

Henriksen, H.J., Barlebo, H.C., 2008. Reflections on the use of Bayesian belief networks for adaptive management. Journal of Environmental Management 88(4), 1025-1036.

Hinneburg, A., Keim, D.A., 1998. An efficient approach to clustering in large multimedia databases with noise, KDD, pp. 58-65.

Holt, J., Mushobozi, W., Day, R., Knight, J., Kimani, M., Njuki, J., Musebe, R., 2006. A simple Bayesian network to interpret the accuracy of armyworm outbreak forecasts. Annals of Applied Biology 148(2), 141-146.

Hopmans, J., Minasny, B., Harter, T., 2003. Neural network prediction of soil hydraulic properties of alluvial soils, EGS-AGU-EUG Joint Assembly.

Houghes, G., 1968. On the mean accuracy of statistical pattern recognition. IEEE Trans. Inform. Theory 14(1), 55-63.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A practical guide to support vector classification.

Hudson, G., Wackernagel, H., 1994. Mapping temperature using kriging with external drift: theory and an example from Scotland. International journal of Climatology 14(1), 77-91.

Irmak, A., Jones, J., Batchelor, W., Irmak, S., Boote, K., Paz, J., 2006. Artificial neural network model as a data analysis tool in precision farming. Transactions of the ASABE 49(6), 2027-2037.

Iyer, M.S., Rhinehart, R.R., 1999. A method to determine the required number of neural-network training repetitions. IEEE Transactions on Neural Networks 10(2), 427-432.

Jafarzadeh, A., Pal, M., Servati, M., FazeliFard, M., Ghorbani, M., 2016. Comparative analysis of support vector machine and artificial neural network models for soil cation exchange capacity prediction. International journal of environmental science and technology 13(1), 87-96.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters 31(8), 651-666.

Jenny, H., 1941. Factors ofSoil Formation, A System of Quantitative Pedology. McGraw-Hill.

Karimi, Y., Prasher, S., Madani, A., Kim, S., 2008. Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. Canadian Biosystems Engineering 50(7), 13-20.

Kassambara, A., 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning, 1. STHDA.

Kaundal, R., Kapoor, A.S., Raghava, G.P., 2006. Machine learning techniques in disease forecasting: a case study on rice blast prediction. BMC bioinformatics 7(1), 485.

Khaki, S., Wang, L., 2019. Crop Yield Prediction Using Deep Neural Networks. arXiv preprint arXiv:1902.02860.

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. European Journal of Soil Science 67(3), 276-284.

Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H., Shamshirband, S., Ab Hamid, S.H., 2015. Developing a fuzzy clustering model for better energy use in farm management systems. Renewable and Sustainable Energy Reviews 48, 27-34.

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.

Kleynhans, T., Montanaro, M., Gerace, A., Kanan, C., 2017. Predicting Top-of-Atmosphere Thermal Radiance Using MERRA-2 Atmospheric Data with Deep Learning. Remote Sensing 9(11), 1133.

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

Koekkoek, E., Booltink, H., 1999. Neural network models to predict soil water retention. European Journal of Soil Science 50(3), 489-495.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, Ijcai. Montreal, Canada, pp. 1137-1145.

Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering 160, 3-24.

Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154(3-4), 340-347.

Kristensen, K., Rasmussen, I.A., 2002. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. Computers and Electronics in Agriculture 33(3), 197-217.

Kubat, M., 2015. An Introduction to Machine Learning. Springer, Place of publication not identified.

Kusumo, B., 2018. In Situ Measurement of Soil Carbon with Depth using Near Infrared (NIR) Spectroscopy, IOP Conference Series: Materials Science and Engineering. IOP Publishing, pp. 012235.

Kusumo, B.H., Hedley, C., Hedley, M., Hueni, A., Tuohy, M., Arnold, G., 2008. The use of diffuse reflectance spectroscopy for in situ carbon and nitrogen analysis of pastoral soils. Soil Research 46(7), 623-635.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296-311.

Lamorski, K., Pachepsky, Y., Sławiński, C., Walczak, R., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Science Society of America Journal 72(5), 1243-1247.

LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., Sackinger, E., 1995. Comparison of learning algorithms for handwritten digit recognition, International conference on artificial neural networks. Perth, Australia, pp. 53-60.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis & Prevention 40(4), 1611-1618.

Li, Y., Shi, Z., Li, F., Li, H.-Y., 2007. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. Computers and Electronics in Agriculture 56(2), 174-186.

Liao, S.-H., Chu, P.-H., Hsiao, P.-Y., 2012. Data mining techniques and applications–A decade review from 2000 to 2011. Expert systems with applications 39(12), 11303-11311.

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

Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment 164, 324-333.

Lobry de Bruyn, L., 2019. Learning opportunities: Understanding farmers' soil testing practice through workshop activities to improve extension support for soil health management. Soil Use and Management.

Lobry de Bruyn, L., Andrews, S., 2016. Are Australian and United States farmers using soil information for soil health management? Sustainability 8(4), 304.

Malone, B.P., Odgers, N.P., Stockmann, U., Minasny, B., McBratney, A.B., 2018. Digital mapping of soil classes and continuous soil properties. Pedometrics, 373-413.

McBratney, A., de Gruijter, J., 1992. A continuum approach to soil classification by modified fuzzy k ‐ means with extragrades. European Journal of Soil Science 43(1), 159-175.

McBratney, A., Pringle, M., 1999. Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. Precision Agriculture 1(2), 125-152.

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

McBratney, A., Webster, R., 1986. Choosing functions for semi ‐ variograms of soil properties and fitting them to sampling estimates. Journal of soil Science 37(4), 617-639.

McBratney, A.B., Odeh, I.O., 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77(2-4), 85-113.

McBratney, A.B., Odeh, I.O., Bishop, T.F., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97(3-4), 293-327.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89(1-2), 67-94.

Mehta, P., Shah, H., Kori, V., Vikani, V., Shukla, S., Shenoy, M., 2015. Survey of unsupervised machine learning algorithms on precision agricultural data, Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on. IEEE, pp. 1-8.

Merdun, H., Çınar, Ö., Meral, R., Apan, M., 2006. Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. Soil and Tillage Research 90(1-2), 108-116.

Milenova, B.L., Campos, M.M., 2002. O-cluster: Scalable clustering of large high dimensional data sets, 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE, pp. 290-297.

Minasny, B., Hopmans, J., Harter, T., Eching, S., Tuli, A., Denton, M., 2004. Neural networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data. Soil Science Society of America Journal 68(2), 417-429.

Minasny, B., McBratney, A., 2002. The neuro-m method for fitting neural network parametric pedotransfer functions. Soil Science Society of America Journal 66(2), 352-361.

Minasny, B., McBratney, A., 2010. Conditioned Latin hypercube sampling for calibrating soil sensor data to soil properties, Proximal soil sensing. Springer, pp. 111-119.

Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma 140(4), 324-336.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301-311.

Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93(3-4), 225-253.

Mohanty, B.P., Cosh, M.H., Lakshmi, V., Montzka, C., 2017. Soil moisture remote sensing: State-of-the-science. Vadose Zone Journal 16(1).

Momma, M., Bennett, K.P., 2002. A pattern search method for model selection of support vector regression, Proceedings of the 2002 SIAM International Conference on Data Mining. SIAM, pp. 261-274.

Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., Mukhopadhyay, A., 2017. Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data. The Egyptian Journal of Remote Sensing and Space Science 20(1), 61-70.

Moran, M.S., Inoue, Y., Barnes, E., 1997. Opportunities and limitations for image-based remote sensing in precision crop management. Remote sensing of Environment 61(3), 319-346.

Mulla, D.J., 2013. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosystems engineering 114(4), 358-371.

Newlands, N.K., Townley-Smith, L., 2010. Predicting energy crop yield using bayesian networks, Proceedings of the Fifth IASTED International Conference, pp. 014-106.

Niang, M.A., Nolin, M.C., Jégo, G., Perron, I., 2014. Digital Mapping of soil texture using RADARSAT-2 polarimetric synthetic aperture radar data. Soil Science Society of America Journal 78(2), 673-684.

Niedbała, G., 2019. Simple model based on artificial neural network for early prediction and simulation winter rapeseed yield. Journal of integrative agriculture 18(1), 54-61.

Odeh, I., Chittleborough, D., McBratney, A., 1992. Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Science Society of America Journal 56(2), 505-516.

Odeh, I.O., McBratney, A., Chittleborough, D., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67(3-4), 215-226.

Odeh, I.O., McBratney, A.B., 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. Geoderma 97(3-4), 237-254.

Odgers, N.P., McBratney, A.B., Minasny, B., 2015. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma 237, 190-198.

Oldfield, E.E., Bradford, M.A., Wood, S.A., 2019. Global meta-analysis of the relationship between soil organic matter and crop yields. Soil 5(1), 15-32.

Oliver, M., Webster, R., 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. Catena 113, 56-69.

Orton, T.G., Mallawaarachchi, T., Pringle, M.J., Menzies, N.W., Dalal, R.C., Kopittke, P.M., Searle, R., Hochman, Z., Dang, Y.P., 2018. Quantifying the economic impact of soil constraints on Australian agriculture: A case‐study of wheat. Land Degradation & Development 29(11), 3866-3875.

Pachepsky, Y.A., Timlin, D., Varallyay, G., 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Science Society of America Journal 60(3), 727-733.

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

Park, S., Hwang, C., Vlek, P., 2005. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. Agricultural Systems 85(1), 59-81.

Park, Y.R., Murray, T.J., Chen, C., 1996. Predicting sun spots using a layered perceptron neural network. IEEE Transactions on Neural Networks 7(2), 501-505.

Peng, J., Loew, A., Merlin, O., Verhoest, N.E., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. Reviews of Geophysics 55(2), 341-366.

Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Pham, D.T., Dimov, S.S., Nguyen, C.D., 2005. Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 219(1), 103-119.

Pierce, F., Anderson, N., Colvin, T., Schueller, J., Humburg, D., McLaughlin, N., Sadler, E., 1997. Yield mapping. The state of site-specific management for agriculture.

Pollino, C.A., Henderson, C., 2010. Bayesian networks: A guide for their application in natural resource management and policy. Landscape Logic, Technical Report 14.

Prabhakar, M., Prasad, Y., Rao, M.N., 2012. Remote sensing of biotic stress in crop plants and its applications for pest management, Crop stress and its management: Perspectives and strategies. Springer, pp. 517-545.

Pringle, M., McBratney, A., Whelan, B., Taylor, J., 2003. A preliminary approach to assessing the opportunity for site-specific crop management in a field, using yield monitor data. Agricultural Systems 76(1), 273-292.

Qiao, D., Shi, H., Pang, H., Qi, X., Plauborg, F., 2010a. Estimating plant root water uptake using a neural network approach. Agricultural water management 98(2), 251-260.

Qiao, D.M., Shi, H.B., Pang, H.B., Qi, X.B., Plauborg, F., 2010b. Estimating plant root water uptake using a neural network approach. Agricultural Water Management 98(2), 251-260.

Ramesh, V., Ramar, K., Babu, S., 2013. Parallel K-Means Algorithm on Agricultural Databases. IJCSI International Journal of Computer Science Issues 10(1), 1694-0814.

Rengasamy, P., 2002. Transient salinity and subsoil constraints to dryland farming in Australian sodic soils: an overview. Australian Journal of Experimental Agriculture 42(3), 351-361.

Robertson, D., Wang, Q.J., 2004. Bayesian networks for decision analyses — an application to irrigation system selection. Australian Journal of Experimental Agriculture 44(2), 145-150.

Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. Science 344(6191), 1492-1496.

Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., Plümer, L., 2010. Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance. Computers and Electronics in Agriculture 74(1), 91-99.

Ruß, G., Kruse, R., 2011. Exploratory hierarchical clustering for management zone delineation in precision agriculture, Industrial Conference on Data Mining. Springer, pp. 161-173.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T., 2016. Meta-learning with memory-augmented neural networks, International conference on machine learning, pp. 1842-1850.

Sarani, F., Ahangar, A.G., Shabani, A., 2016. Predicting ESP and SAR by artificial neural network and regression models using soil pH and EC data (Miankangi Region, Sistan and Baluchestan Province, Iran). Archives of Agronomy and Soil Science 62(1), 127-138.

Schaap, M.G., Bouten, W., 1996. Modeling water retention curves of sandy soils using neural networks. Water Resources Research 32(10), 3033-3040.

Schaap, M.G., Leij, F.J., Van Genuchten, M.T., 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Science Society of America Journal 62(4), 847-855.

Schoier, G., Borruso, G., 2004. A clustering method for large spatial databases, International Conference on Computational Science and Its Applications. Springer, pp. 1089-1095.

Seif, G., 2018. The 5 Clustering Algorithms Data Scientists Need to Know.

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

Sibson, R., 1981. A brief description of natural neighbour interpolation. Interpreting multivariate data.

Sirjacobs, D., Hanquet, B., Lebeau, F., Destain, M.-F., 2002. On-line soil mechanical resistance mapping and correlation with soil physical properties for precision agriculture. Soil and Tillage Research 64(3-4), 231-242.

Smith, C.S., Howes, A.L., Price, B., McAlpine, C.A., 2007. Using a Bayesian belief network to predict suitable habitat of an endangered mammal–The Julia Creek dunnart (Sminthopsis douglasi). Biological Conservation 139(3), 333-347.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and computing 14(3), 199-222.

Stafford, J., Ambler, B., Lark, R., Catt, J., 1996. Mapping and interpreting the yield variation in cereal crops. Computers and Electronics in Agriculture 14(2-3), 101-119.

Steeneveld, W., van der Gaag, L.C., Barkema, H.W., Hogeveen, H., 2010. Simplify the interpretation of alert lists for clinical mastitis in automatic milking systems. Computers and electronics in agriculture 71(1), 50-56.

Stephanie, 2016. What is Hierarchical Clustering?

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

Suykens, J.A., Vandewalle, J., 2000. Recurrent least squares support vector machines. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 47(7), 1109-1114.

Tamari, S., Wösten, J., Ruiz-Suarez, J., 1996. Testing an artificial neural network for predicting soil hydraulic conductivity. Soil Science Society of America Journal 60(6), 1732-1741.

Tari, F., 1996. A Bayesian Network for predicting yield response of winter wheat to fungicide programmes. Computers and Electronics in Agriculture 15(2), 111-121.

Thomasson, J.A., Baillie, C.P., Antille, D.L., Lobsey, C.R., McCarthy, C.L., 2019. Autonomous Technologies in Agricultural Equipment: A Review of the State of the Art. American Society of Agricultural and Biological Engineers.

Tilling, A.K., O'Leary, G.J., Ferwerda, J.G., Jones, S.D., Fitzgerald, G.J., Rodriguez, D., Belford, R., 2007. Remote sensing of nitrogen and water stress in wheat. Field Crops Research 104(1-3), 77-85.

Tranter, G., Minasny, B., McBratney, A., 2010. Estimating Pedotransfer Function Prediction Limits Using Fuzzy k-Means with Extragrades All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. Soil Science Society of America Journal 74(6), 1967-1975.

Troldborg, M., Aalders, I., Towers, W., Hallett, P.D., McKenzie, B.M., Bengough, A.G., Lilly, A., Ball, B.C., Hough, R.L., 2013. Application of Bayesian Belief Networks to quantify and map areas at risk to soil threats: Using soil compaction as an example. Soil and Tillage Research 132, 56-68.

Tsiropoulos, Z., Fountas, S., Gemtos, T., Gravalos, I., Paraforos, D., 2013. Management information system for spatial analysis of tractor-implement draft forces, Precision agriculture'13. Springer, pp. 349-356.

Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology 49(11), 1225-1231.

Twarakavi, N.K., Šimůnek, J., Schaap, M., 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. Soil Science Society of America Journal 73(5), 1443-1452.

Üstün, B., Melssen, W.J., Oudenhuijzen, M., Buydens, L.M.C., 2005. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. Analytica Chimica Acta 544(1), 292-305.

Van Bergeijk, J., Goense, D., 1996. Soil tillage resistance as tool to map soil type differences. Precision Agriculture (precisionagricu3), 605-616.

Van Der Gaag, L.C., Bolt, J., Loeffen, W., Elbers, A., 2010. Modelling patterns of evidence in Bayesian networks: a case-study in classical swine fever, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, pp. 675-684.

Van Egmond, F., Loonstra, E., Limburg, J., 2010. Gamma ray sensor for topsoil mapping: The Mole, Proximal Soil Sensing. Springer, pp. 323-332.

Vapnik, V., 1995. The nature of statistical learning theory. J. Wiley & Sons, New York.

Varis, O., 1997. Bayesian decision analysis for environmental and resource management. Environmental Modelling & Software 12(2), 177-185.

Viscarra Rossel, R.A., Adamchuk, V., Sudduth, K., McKenzie, N., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time, Advances in agronomy. Elsevier, pp. 243-291.

Viscarra Rossel, R.A., McBratney, A., 1998. Laboratory evaluation of a proximal sensing technique for simultaneous measurement of soil clay and water content. Geoderma 85(1), 19-39.

Viscarra Rossel, R.A., McBratney, A.B., Minasny, B., 2010. Proximal soil sensing. Springer Science & Business Media.

Waiser, T.H., Morgan, C.L., Brown, D.J., Hallmark, C.T., 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Science Society of America Journal 71(2), 389-396.

Wani, M.A., Bhat, F.A., Afzal, S., Khan, A.I., 2019. Advances in Deep Learning. Springer.

Warrick, A.W., 2001. Soil physics companion. CRC press.

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

Wendelberger, J.G., 1981. The Computation of Laplacian Smoothing Splines with Examples, WISCONSIN UNIV-MADISON DEPT OF STATISTICS.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52, 394-403.

Wessels, L.F., Barnard, E., 1992. Avoiding false local minima by proper initialization of connections. IEEE Transactions on Neural Networks 3(6), 899-905.

Whelan, B., McBratney, A., 2000. The "null hypothesis" of precision agriculture management. Precision Agriculture 2(3), 265-279.

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

White, H., 1989. Learning in artificial neural networks: A statistical perspective. Neural computation 1(4), 425-464.

Widrow, B., Lehr, M.A., 1990. 30 years of adaptive neural networks: perceptron, Madaline, and backpropagation. Proceedings of the IEEE 78(9), 1415-1442.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wong, M., Wittwer, K., Oliver, Y., Robertson, M., 2010. Use of EM38 and gamma ray spectrometry as complementary sensors for high-resolution soil property mapping, Proximal soil sensing. Springer, pp. 343-349.

Wösten, J., Finke, P., Jansen, M., 1995. Comparison of class and continuous pedotransfer functions to generate soil hydraulic characteristics. Geoderma 66(3-4), 227-237.

Xu, X., Ester, M., Kriegel, H.-P., Sander, J., 1998. A distribution-based clustering algorithm for mining in large spatial databases, Proceedings 14th International Conference on Data Engineering. IEEE, pp. 324-331.

Yan, L., Zhou, S., Feng, L., 2007. Delineation of Site-Specific Management Zones Based on Temporal and Spatial Variability of Soil Electrical Conductivity1. Pedosphere 17(2), 156-164.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A., Hann, S., Burt, J.E., Qi, F., 2011. Updating conventional soil maps through digital soil mapping. Soil Science Society of America Journal 75(3), 1044-1053.

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

Zhang, Y., Biswas, A., Ji, W., Adamchuk, V.I., 2017. Depth-specific prediction of soil properties in situ using vis-NIR spectroscopy. Soil Science Society of America Journal 81(5), 993-1004.

Zhao, G., Bryan, B.A., King, D., Luo, Z., Wang, E., Bende-Michl, U., Song, X., Yu, Q., 2013. Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing. Environmental Modelling & Software 41, 231-238.

Zhou, A., Zhou, S., Cao, J., Fan, Y., Hu, Y., 2000. Approaches for scaling DBSCAN algorithm to large spatial databases. Journal of computer science and technology 15(6), 509-526.

3. Description of experimental site and general soil analytical methods employed in this work

3.1. Introduction

This study aims to identify a minimum soils dataset required to understand soil function at a spatial scale using directly measured and proximally sensed data. This requires an intensive, directly measured spatial dataset to be collected and analysed; an exercise that has not been undertaken in Australian agriculture to the level required to accurately assess this question. An intensive field-based sampling regime was conducted and samples analysed in a laboratory for structural and chemical properties. Machine learning (ML) techniques were used to investigate spatial soil function and its relationship with crop yield variability at the field scale, a key metric for soil performance.

3.2. Field Selection Criteria

The investigated site required an appropriate level of crop yield maps in order to appropriately asses the field variability and its relationship with soil function. At this point it is noted that the site should have had two wheat crops planted and harvested throughout the lifetime of the project. However, drought conditions resulted in the crops not being planted, which did limit the available yield data, but was beyond the control of the project.

The site needed to be representative of a key dryland cropping region to ensure results were extendable for future work. Crop rotations at the site were consistent with that commonly practiced throughout Central New South Whales (winter cereal and pulse rotation of wheat/barley/chickpea and canola). In additional to this, the soil constraints at the site were largely representative of that managed in the (i.e. mainly sodicity and compaction, with some surface acidity). Controlled Traffic Farming (CTF) practices were pertinent in reducing the influence of previous random traffic which cannot be accounted for. Key agricultural consultants across New South Wales and Queensland were engaged in locating an appropriate site. The selected site was also representative of farming systems in Central New South Whales

3.3. Sampling District

The selected site is located within the Warren district of the Macquarie Valley in central NSW, as shown in Figure 3.2. The Macquarie Valley is described as Grassland – Hot (persistently dry) using the Climate Classification of Australia (Bureau of Meteorology, 2018).

The mean climate data for the sampling district is displayed in Table 3.1 and Figure 3.1. The district consists mainly of dryland cropping, dryland grazing and mixed farming management, with some concentrated irrigated cropping following the Macquarie River. Dryland cropping is dominated with the production of winter cereals, including barley, wheat, chickpea and canola. Chromosols, Dermosols and Kandosols are the dominant soil type within the immediate Bundemar region of the experimental site (McKenzie, 1992), as classified using the Australian Soil Classification System (Isbell, .)

Table 3.1 Weather statistics for the experimental site

Valley	*Region*	*Annual Rainfall*	*Mean number of days with rainfall >1mm*	*Annual Temperature (°C)*		
				Minimum	*Maximum*	*Mean number of days > 30°C*
Macquarie	Bundemar	413	49	10	25	93

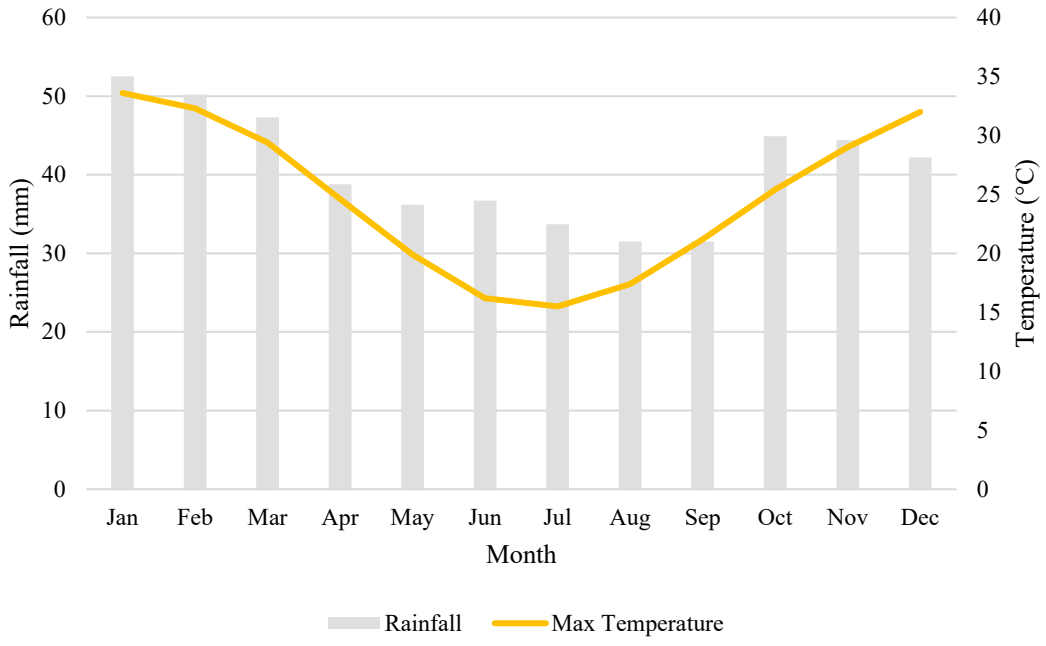


Figure 3.1 Climate statistics for the Bundemar region

3.4. "Fiona Downs" Bundemar, Warren

3.4.1. History and on-field operations

The investigation site is situated in the Bundemar region, approximately 30km south-east of Warren, NSW and 25km north-east of Trangie, NSW (GR 31°49'40.49" S 148°06'44.56"E). The 108 ha site is located on a property that is currently corporately owned and managed, and has been since 2000. The site has been managed solely for dryland winter

cereals since ownership (Table 3.2). Prior to this it is thought that the site was grazed. The site has been under CTF management since 2009 using a 12 m farming system as well as zero-till practices since 2010. Historically, the site has possessed a high degree of crop yield variability that is evident in the available yield maps (Figure 3.3). It is suggested that this is due to spatial variations in soil condition and soil function across the site (Precision Cropping Technologies, pers comms, Mar 2017). In-field operations used for winter cereals are generally consistent with that of the region and have not included variable rate application of seed, fertiliser, chemical or soil ameliorants to this date, despite the known variations that exist.

Table 3.2. Total rainfall for each growing season taken from November of the previous year to October of the cropped year

Year	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*
Rotation	Wheat	Chickpea	Wheat	Wheat	Chickpea*	Wheat*
Rainfall	381	380.3	500.8	691	307.1	333.2

* Unseasonably dry conditions resulted in failed crop

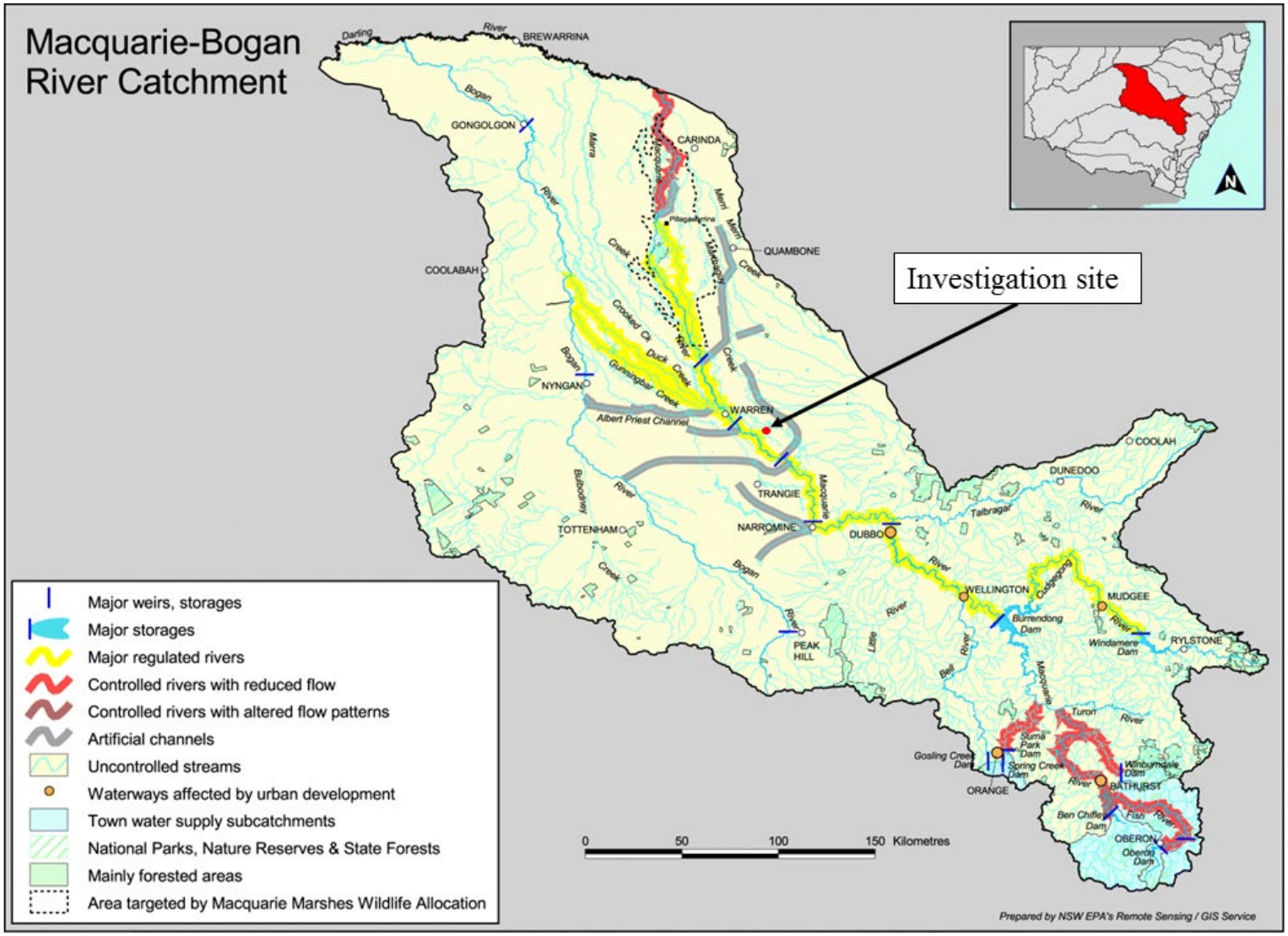
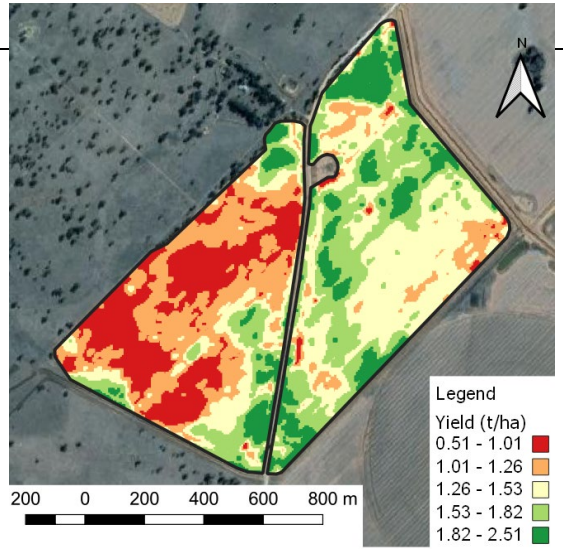


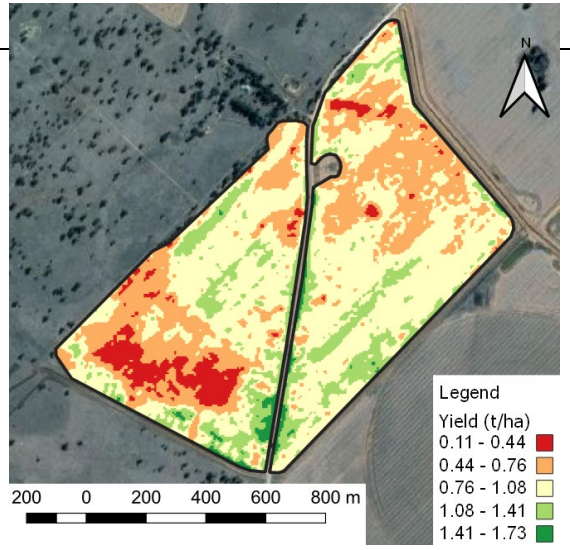
Figure 3.2. The Macquarie Valley. Source: https://www.environment.nsw.gov.au/ieo/MacquarieBogan/maplg.htm

Table 3.3. Summary statistics of measured soil properties at the investigation site

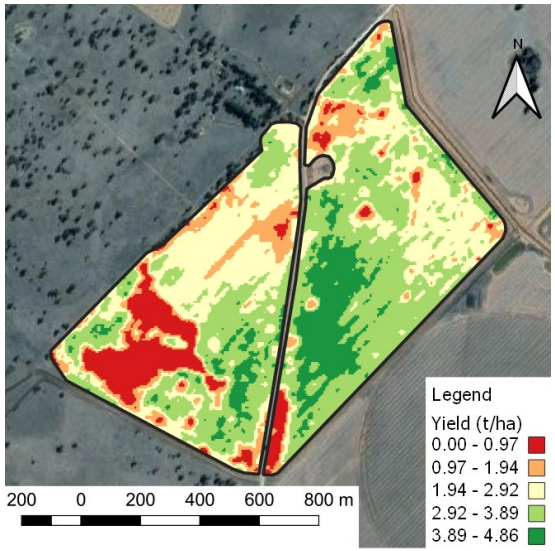
Depth	*Statistic*	*pH*	*EC (ds/m)*	*Clay (%)*	*Silt (%)*	*Sand (%)*	*Moisture (θ_g)*	*BD (m/m^3)*	*Na ([cmol(+)/kg])*	*Mg ([cmol(+)/kg])*	*K ([cmol(+)/kg])*	*Ca ([cmol(+)/kg])*	*CEC ([cmol(+)/kg])*	*EMGP*	*EDP*	*ESP*
0–10 cm	Min	5.27	0.04	3.75	15	10.0	2.45	1.18	0.2	0.69	0.43	3.96	5.78	12.0	3.51	0.03
	Max	9.15	0.29	71.3	52.5	65.0	24.6	1.83	4.73	9.90	2.66	30.9	38.3	38.6	23.4	20.9
	Average	6.58	0.1	39.5	28.9	31.6	8.69	1.47	0.62	4.49	1.42	10.0	16.6	27.4	9.22	4.01
	SD	0.64	0.04	10.3	6.52	8.12	2.22	0.11	0.53	1.70	0.39	4.95	6.39	5.64	3.1	3.17
10–20 cm	Min	5.98	0.03	8.75	3.75	13.8	4.93	1.37	0.01	1.04	0.21	4.64	7.64	10.4	1.7	0.13
	Max	9.23	0.37	72.5	45	63.8	30.8	1.84	5.47	11.1	3.51	29.4	39.2	46.2	27.1	26.2
	Average	7.52	0.09	47.8	25.6	26.6	12.8	1.61	1.24	6.88	0.90	14.3	23.3	29.5	7.64	5.32
	SD	0.68	0.04	9.24	5.85	7.64	2.62	0.08	0.93	1.96	0.40	4.58	6.16	5.06	3.37	3.75
20–40 cm	Min	6.55	0.04	20	6.25	8.75	6.55	0.80	0.01	2.48	0.20	6.27	10.1	11.7	2.35	0.05
	Max	9.45	0.49	73.8	42.5	55.0	57.7	1.85	9.31	22.1	2.98	39.7	66.5	41.8	31.3	30.3
	Average	8.23	0.15	50.4	25.6	24.0	15.1	1.64	2.13	8.58	0.73	16.8	28.2	30.5	8.85	7.36
	SD	0.57	0.07	6.98	5.85	6.16	3.43	0.08	1.45	1.94	0.36	3.85	5.46	4.3	4.43	4.69
40–60 cm	Min	5.98	0.06	20	3.75	11.3	8.81	0.98	0.03	2.48	0.2	6.74	11.1	19.9	2.31	0.14
	Max	9.65	1.65	67.5	47.5	51.3	51.1	1.91	10.4	14.2	2.39	27.6	42.0	44.1	37.1	34
	Average	8.72	0.24	50.1	25.8	24.1	15.0	1.68	3.20	9.19	0.59	16.6	29.6	31.2	11.7	10.5
	SD	0.56	0.15	6.92	6.28	6.18	3.31	0.07	1.91	1.68	0.31	2.98	4.53	3.71	5.59	5.9



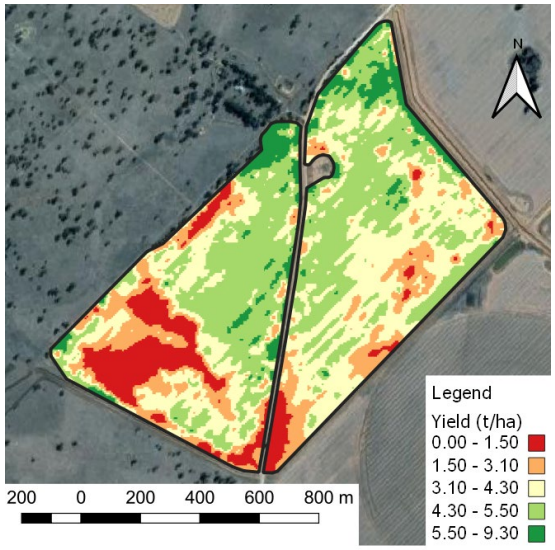
2013 Wheat Yield



2014 Chickpea Yield



2015 Wheat Yield



2016 Wheat Yield

Figure 3.3. Historic yield maps for the investigation site

3.4.2. Soil profile characterisation

Summary statistics for measured soil properties at the site are presented in Table 3.3, which exhibit a high degree of spatial variability. The investigated site occupies some acidic conditions at the surface, but is generally alkaline to 60 cm. Soil textures range from clay loams at the surface to medium clays in the subsurface layers. A discrete section within the north-east area of the site exhibited a sandy clay loam class. Low EC values were observed throughout the profile, suggesting that the site was not saline. BD was generally high (Hazelton and Murphy, 2016) in all surface layers and increased to depth; it is likely that the site is compaction constrained to some degree. ESP at the site varied substantially in the x, y and z planes, with non-sodic to highly sodic conditions being observed within each depth. Both the magnitude and variability of ESP increased with depth.

3.5. Methods

3.5.1. Soil Sampling

Sampling was undertaken in June of 2017, 3 weeks after the site was sown to chickpea. and the sampling density consisted of a 60 m grid sampling pattern to fit within the 12m farming system that was present (Figure 3.2). This resulted in a total of 300 sampling locations for the site. At each sampling location, two intact soil cores were extracted to 60 cm; one for chemical analysis and the other for BD and moisture measurements. The two cores were extracted at a distance of approximately 10 cm in the direction of field traffic. This was achieved by moving the sampling vehicle 10 cm after the first extraction. Cores were extracted using a core sleeve with an internal cutting tip diameter (ID) of 43 mm on a utility-mounted hydraulic coring apparatus with optional jackhammer action; a minimal amount of synthetic lubricant was used for each core to aid in removal of the cores from the sleeve, while also minimising sample contamination. As the sampling method used a hammering action, soil core length was measured and the depth of the hole confirmed to ensure compaction had not occurred during sampling; this method did not cause compaction of samples. Each core was sectioned into the key agronomic depths of 0–10 cm, 10–20 cm, 20–40 cm and 40–60 cm, assumed to define the zone of bulk rooting density, and stored in sampling bags. A total of 1200 samples were collected from the site. Following the 2-week sampling period, the samples were immediately returned to the laboratory, oven dried at 40 °C and

were crushed and sieved to a stabilised 2 mm fraction (for chemical analysis). A proportion of aggregates from each sample were kept aside for further analysis.

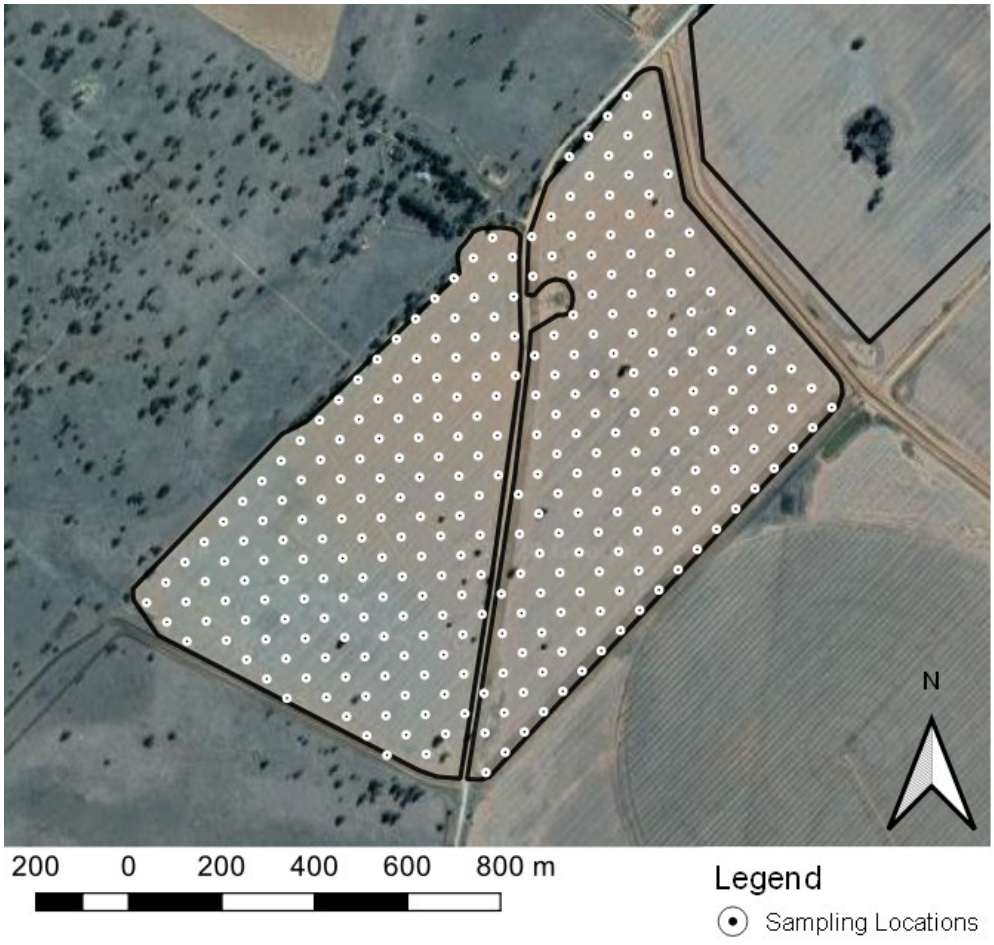


Figure 3.4. Sampling locations of the experimental site

Figure 3.5. Soil coring apparatus

3.5.2. Particle Size Analysis

Particle size distribution was determined using the hydrometer method described by Gee and Bauder (1986). Forty grams of stabilised soil were measured, and the exact weight recorded, and placed into 250 mL plastic bottles. Each bottle was subsequently filled with 50 mL of 10 % sodium hexa-metaphosphate (HMP) (w/w), 5 mL of 0.6 M of sodium Hydroxide and 100 mL of distilled water. Afterwards, each bottle was capped and placed in an end-over-end shaker for 24 h. The entirety of the soil and liquid was transferred to a 1.0 L measuring cylinder and made to 1.0 L with distilled water before being homogenised with a baffle rod. Each solution was allowed to stand for the appropriate period of time before hydrometer readings were taken; clay and silt were measured at 5 min, while clay was measured after 5 hours. Clay, silt and sand percentages were subsequently determined.

3.5.3. Soil pH

Soil pH was determined in 1:5 soil:water extract consistent with (Rayment and Lyons, 2011) – method 4A1. Eight grams of soil were placed into 50 mL centrifuge tubes and 40 mL of

deionised water added. Each tube was placed in an end-over-end shaker for 1 h and centrifuged at 3500 rpm for 30 min to settle out soil particles from the solution. Soil pH was determined using Radiometer Analytical^{TM} pH meter (TIM845) with an automatic temperature calibration.

3.5.4. Soil electrical conductivity (EC)

EC was determined, consistent with (Rayment and Lyons, 2011)using the same 1:5 soil:water extract used for pH measurement. Measurements were taken using an Orion Star^{TM} conductivity meter (A212) with automatic temperature calibration.

3.5.5. Exchangeable cations

The major exchangeable cations (Na^+, Mg^{2+}, K^+, Ca^{2+}) were measured using two different methods in (Rayment and Lyons, 2011) — 3.5.5.A and 3.5.5.B— dependent on whether a sample's EC reading warranted pre-treatment for soluble salts (>500 dS/m). The following reagents and flame solutions were used:

- Reagent 1: 1 M ammonium chloride (NH_4CL), adjusted to a pH of 7 and 8.5 with ammonium in order to aid in the suppression of carbonate dissolution in soils with a high pH (>7.8).
- Reagent 2: 60% Aqueous Ethanol (w/w). A volume of 665 mL of 96% ethanol were made to 1 L with deionised water.
- Flame Solution: 1667 mg/L strontium chloride (Sr)
-

3.5.5.A. Exchangeable bases in 1 M ammonium chloride, no pre-treatment for soluble salts

The method employed is adapted from 15A1 (Rayment and Lyons, 2011), and was applied to samples with EC<500 dS/m, employing the use of one reagent and a flame solution (for analysis by atomic absorption spectrophotometry; AAS).

Two grams of stabilised soil were measured and placed into 50 mL centrifuge tubes. Samples were separated according to pH (<7.8 and ≥7.8) to ensure to the appropriate pH adjusted

reagent was used (pH 7 reagent 1 used for samples with 1:5 pH<7.8 and pH 8.5 reagent 1 used for samples with 1:5 pH≥7.8). A volume of 40 mL of reagent was added. Samples were placed in an end-over-end shaker for 1 h and centrifuged at 3500 rpm for 10 min to settle soil particles. A 10 mL of the soil solution was immediately transferred into 15 mL falcon tubes to ensure further exchange did not occur. Samples were diluted with the flame solution to 1:10, giving a final dilution of 1:200. Standards were prepared using the same matrix.

3.5.5.B. Exchangeable bases in 1 M ammonium chloride, pre-treatment for soluble salts

The method employed was adapted from 15A1 in Rayment and Lyons (2011), and applies to samples with EC>500 dS/m. The same extracting reagent and flame solution described in 3.5.5.A were used, with the addition of reagent 2.

Two grams of stabilised soil were measured and placed into 50 mL centrifuge tubes and 25 mL of reagent 2 added (for samples >500 dS/m). Samples were sealed and placed in an end-over-end shaker for 30 mins before being centrifuged at 3500 rpm for 10 minutes to remove the supernatant solution. Reagent 2 was drained from each sample, ensuring no clay particles were lost and left to drain upside down on a piece of absorbent paper for 5 min to remove excess solvent. This entire process was repeated a second time before method 3.5.5.A was employed to extract the specific exchangeable cations.

3.5.6. Moisture content

Soil samples were weighed to determine the field wet mass of each sample before being oven dried at 105 °C for 72 h, and weighed again to determine the oven dry mass of the samples. By difference of the wet and oven dry mass, the gravimetric moisture content was determined. The volumetric moisture content was subsequently calculated one BD of the sample was obtained.

3.5.7. Bulk density

BD was measured from in-tact soil cores that were carefully segmented into the prescribed depth intervals (section 3.5.1), and the cross-sectional area of the soil core cutting tip assumed as the soil core area. Using these dimensions, the volume of each sample was calculated (146.91 cm^3 for 10 cm samples and 293.83 cm^3 for 20 cm samples). The same oven dried sample weights

described in 3.5.6. were used to calculate the BD of each sample and reported as mass per volume (g cm^3). The weight of the sample bag was accounted for during the calculation.

3.5.8. Spatial data kriging

Analysis of the data required all data points to be geospatially aligned on a common grid to ensure continuity. All data layers were kriged to a common grid using the *automap* package in the R programming environment (Hiemstra and Hiemstra, 2013). A grid size of 6x6 m was selected to appropriately fit into both the scale of the farming system (12 m) and the sampling regime (60 m). The kriging process resulted in a total of 29,978 interpolated data points for each data layer for the site. Variogram parameters for pH and ESP at the 4 investigated depths are presented in Table 3.4.

Table 3.4 Parameters of the fitted pH and ESP variograms for the benchmark dataset.

	Depth	Model	Nugget	Sill	Range	Kappa
ESP	0–10 cm	Ste	.78	11	403	.2
	10–20 cm	Ste	4.7	17	820	0.2
	20–40 cm	Ste	0.66	22	257	0.2
	40–60 cm	Ste	4.1	48	771	0.3
pH	0–10 cm	Ste	0.14	0.42	361	0.9
	10–20 cm	Ste	0.16	1	4374	0.2
	20–40 cm	Ste	0.03	0.33	149	0.2
	40–60 cm	Ste	0.09	0.33	277	0.2

3.6. References

Gee, G., Bauder, J., 1986. Particle-size analysis. In ‘Methods of soil analysis. Part 1. Physical and mineralogical methods’.(Ed. A Klute) pp. 383–411. Soil Science Society of America: Madison, WI, USA.

Hazelton, P., Murphy, B., 2016. Interpreting soil test results: What do all the numbers mean? CSIRO publishing.

Hiemstra, P., Hiemstra, M.P., 2013. Package ‘automap’. compare 105, 10.

Isbell, R., 1996. The Australian Soil Classification CSIRO Publishing. Collingwood, Australia.

McKenzie, N.J., 1992. Soils of the Lower Macquarie Valley, New South Wales. CSIRO Division of Soils.

Rayment, G.E., Lyons, D.J., 2011. Soil chemical methods: Australasia, 3. CSIRO publishing.

4. Assessing the sensitivity of site-specific lime and gypsum recommendations to soil sampling techniques and spatial density of data collection: A pedometric approach

4.1. Introduction

Site-specific agronomic decisions are often made using limited soil information, due to the perceived cost of soil data acquisition in relation to its perceived usefulness (Bennett and Cattle, 2013; Bennett and Cattle, 2014; Lobry de Bruyn and Andrews, 2016). As such, agricultural advisors typically take a data minimal approach, frequently using surface-based "grab-samples" (i.e. approximate 0–10 cm depth) along a transect and bulking these as a single representation of the field condition. The bulked samples are analysed and subsequently used to diagnose the field requirement for soil management. In doing this the ability to identify the variability of site characteristics is lost, severely limiting the sampling data from which agronomic recommendations are made. Whilst this approach represents current practice in industry, there is limited understanding surrounding the economic and agronomic consequences of decisions made in such a data limited environment. More specifically, there has been limited assessment of the magnitude of error associated with agronomic recommendations at various sampling densities. The financial ramifications of spatially inaccurate agronomic recommendations have the potential to be highly influential on overall farm profitability, as large soil treatment investments are often made on their basis (Bennett et al., 2015a; Bennett et al., 2016). Furthermore, there is an emerging social responsibility for advisors and land owners to ensure appropriate management of farming inputs to match soil conditions (Bennett, 2019; Heath, 2018; Lush, 2018).As such, it is prudent to understand how the spatial nature of constraints affects amendment application regimes, as well as the data requirement to sufficiently manage soils for a soil condition that is both productive and profitable, whilst simultaneously demonstrating the social responsibility of management (Bennett, 2019).

Transect and low density sampling designs commonly take a randomised approach to selection of the sample locations, or at the very least, the transect initiation point and direction of traverse (Pennock et al., 2007). This means the accuracy of these methods is constrained by the inability to capture spatial variance (De Gruijter et al., 2006). This is associated with a large degree of error, the magnitude of which is largely unknown to the land manager. More advanced random sampling is possible in the presence of auxiliary information (e.g. crop yield data, remotely sensed data, proximally sensed data etc.) which involves a clustering approach,

whereby a field is stratified into discrete spatial strata (De Gruijter, 1977). These strata are commonly referred to as *management zones*, from which samples are targeted and recommendations made, on a zone-average basis. The approach evolved to identify and map soil classes across a landscape, with key examples for management zone delineation including Boydell and McBratney (2002), Fu et al. (2010), Li et al. (2007) and Taylor et al. (2007). Clustering for zone-based management is the current accepted standard for variable rate management in commercial precision agriculture (McBratney et al., 2005; Robertson et al., 2012) although the inability to map continuous soil properties significantly reduces the capacity for true precision management.

Fuzzy logic improves traditional clustering techniques by overcoming the limitations of forcing hard boundaries between continuous soil classes (De Gruijter and McBratney, 1988; McBratney and de Gruijter, 1992a). These continuous soil classes are classified by statistically assigning a continuous membership value between each point in the population and the centroids of the defined clusters, depending on the degree of similarity within the auxiliary data (McBratney and Odeh, 1997). Allowing fuzzy boundaries between continuous soil classes also facilitates the measurement of uncertainty between the memberships of observations and their identified soil class. However, fuzzy logic does not provide the capability to continuously map individual soil properties across the landscape, which is a limitation where agricultural management inputs are based on these individual soil characteristics.

Digital soil mapping (DSM) offers a suite of approaches specifically designed to continuously map individual soil properties across a landscape, using either geostatistical or non-interpolation approaches. Geostatistical approaches use only directly measured or sensed soil observations at known locations to provide inference towards the likely conditions at neighbouring locations, via interpolation, without relying on auxiliary information. The following assumption forms the basis of geostatistical DSM approaches:

$$S = f(x, y), s(x + u, y + v) \qquad \text{Equation 4.1}$$

where soil at some location x,y is a function of the geographic coordinates *x,y* and soil at neighbouring locations (*x+u, y+v*) (McBratney et al., 2003). A number of geostatistical approaches exist, namely: i) inverse squared distance interpolation; ii) natural neighbour interpolation (Sibson, 1981); iii) quadric trend surface; iv) Laplacian smoothing splines (Wendelberger, 1981); and, v) ordinary kriging (OK) (Burgess and Webster, 1980a). Of these,

OK has been the most widely adopted interpolation technique in soil science (McBratney et al., 2000).

For a given soil property, OK aims to fit a semi-variogram to the spatial variance within the data, weighting the level of influence observation points have on a predicted location, based on distance. Optimal sampling patterns for spatial interpolation are those at a regular grid (Brus and Heuvelink, 2007; Heuvelink et al., 2006; Vašát et al., 2010), such as triangular or square lattices, however, consideration towards randomisation of the grid is required to ensure the sampling domain is adequately represented in the sampling regime. Whilst optimal sample spacing for grid designs can be determined using the level of precision required for the soil survey (McBratney et al., 1981), the minimum sample density required for adequate variogram estimation as indicated by Webster and Oliver (1992) has generally not been available at the field level within agriculture. Recognising this, McBratney et al. (2003) extended the seminal work of Jenny (1941) to formalise the framework for non-interpolation DSM approaches, referred to as the *scorpan* framework, which allows for continuous mapping of individual soil properties.

Scorpan models are more formally referred to as soil spatial prediction functions (SSPF) (Malone et al., 2018), and aim to fit numerical models between soil observations and scorpan factors (Minasny and McBratney, 2016), under the assumption of environmental correlation:

$$S_a = f(s, c, o, r, p, a, n) + e \qquad \text{Equation 4.2}$$

which assumes any soil property, 'a', at a given location is a function of other soil properties at that location based on: space (s), climate (c), organisms (o), relief (r), parent material (p), age (a), spatial position (n) and auto-correlated errors (e). Possible representations of each factor are provided in Table 4.1. They key benefit to SSPF models is the ability to map soil properties utilising the full data set of environmental covariates that are often collected at a more spatially exhaustive resolution than the discretely sampled soil depth information. In terms of modelling approaches, linear regression, logistic regression, regression trees, neural networks, support vector machines (SVMs), quantile regression forests and random forests have been used for SSPF development (Malone et al., 2018). Importantly, the addition of regression kriging within the model allows for spatially auto-correlated residuals to be added to spatial predictions, which improves model performance (Hudson and Wackernagel, 1994;

Knotters et al., 1995; Odeh and McBratney, 2000) by accounting for model errors spatially. This is referred to as SSPFe.

Table 4.1 Possible representations of the scorpan factors (after Malone et al., 2018)

scorpan factor	*Possible representations*
s	Legacy soil maps, point observations, expert knowledge
c	Temperature and precipitation records
o	Vegetation maps, species and abundance maps, yield maps, land use maps
r	Digital elevation model, terrain attributes
p	Legacy geology maps, gamma radiometric information
a	Weathering indices, geology maps
n	Latitude and longitude or easting and northing, distance from landscape features, distance from roads, distance from point sources of pollution

Calibration sample selection for SSPFe requires consideration towards both the feature space of the environmental covariates and the geographic space (Brus and Heuvelink, 2007; Hengl et al., 2003; Lesch et al., 1995; Müller, 2001). Optimising model training benefits from a large spread of predictor variables (feature space), whereas residual kriging benefits from even distribution across the geographic space. Heuvelink et al. (2006) however discovered that consideration towards optimal sample distribution across the feature space is more important than attempting to optimise model training within the geographic space. On this basis, conditioned Latin hypercube sampling (cLHS) (Minasny and McBratney, 2006b) has become a widely used technique for calibration sample selection in DSM. The aim of cLHS is to select sampling points across the feature space of the environmental covariates by considering their multivariate distributions. Subsequently, it is typically advantageous to select calibration samples towards the extremes of the feature space to ensure appropriate model fit across the range of variables (Minasny and McBratney, 2010). To achieve this, the extremities of the feature space are weighted, increasing their likelihood of selection, which is referred to as (DLHS) and is inspired by the D-optimality criterion of linear regression (John and Draper, 1975). However, Minasny and McBratney (2010) identified that, if the relationship between the covariates and soil properties is not known (i.e. the majority of DSM applications), cLHS outperforms DLHS and should be the method of choice for sample calibration selection.

The development of these pedometric approaches has primarily been driven on the requirement to rapidly deal with large spatial domains and/ or low density data sets for mapping of soil attributes. We contend that there is agronomic merit in producing localised on-farm soil-crop calibration datasets at a magnitude not generally collected in industry, therefore providing

opportunities for DSM approaches at much finer resolutions than currently practiced. The cLHS method provides a method for sample selection for SSPFe calibration, but there has been limited attention provided to the effect of calibration sample density on spatial predictive accuracy. Hence, there is an associated paucity of information concerning the spatial errors in agronomic recommendations at the field-scale. Furthermore, the majority of SSPFe adoption in the literature has focused on developing models at the regional level, so it is unclear how appropriate the non-interpolation approaches are for site-specific calibration within precision agriculture. Therefore, the aim of this work is to compare the field-scale performance of the sampling techniques identified above, at increasing sampling density, with assessment of their sensitivity in the provision of agronomic recommendations. Furthermore, we will assess the level of uncertainty surrounding these methods via random initialisation of the search parameters within each method.

4.2. Materials and Methodologies

4.2.1. Experimental design

A dataset was collected using a 60 m sampling grid to a depth of 60 cm for the 108 ha experimental site, resulting in 300 soil cores with 4 analysis depths (1200 samples in total). This data density was selected as a pragmatic and resource constrained intensity, and used as the baseline of observed variability at the site. Whilst this assumption is clearly incorrect, it was considered reasonable on the basis of being proximal to, if not surpassing, the upper limit of economically feasible sampling density — approximate cost for soil analysis of $833/ha, excluding resources required for fieldwork. Subsequently, this baseline sample density provided the basis on which the identified sampling methods were simulated to select samples from the observed dataset and produce spatial maps and agronomic recommendations accordingly.

The methods investigated were: 1) random transect sampling, 2) zonal sampling, 3) OK, and 4) SSPF regression kriging. Each method was assessed at separate sampling densities of 10, 20, 50, 100, 150, 200, 250, and 300. Each method was applied to create a predicted DSM of pH, ESP, CEC and BD at 6 x 6 m pixels, from which agronomic recommendations were made. These predicted DSMs were spatially compared against the benchmark DSM created from the 300 directly measured soil properties for each depth, using OK. Spatial prediction errors of soil attributes were calculated as:

$$Error_{x,y}^{i} = abs(P_{x,y}^{i} - O_{x,y}^{i})$$ Equation 4.3.

where P and O are the respective predicted and observed value of soil attribute *i* at grid location *x,y*. Root-mean square errors (RMSE) were subsequently calculated to provide an indication on average error, as well as the interquartile ranges of the prediction residuals to provide insight into the range of prediction error at various sampling densities across the methods.

4.2.2. Investigation Site

The investigation site is located within the Warren district of the Macquarie Valley in central NSW, Australia (GR 31°49'40.49" S 148°06'44.56"E). The 108ha dryland site is managed as a 12 m CTF frontage, zero-tillage farming system and is under a winter cropping rotation consisting dominantly of wheat, chickpea and barley. The dominant soil types identified at the site were Kandosols and Dermosols as classified using the Australian Soil Classification System (Isbell 2002). Minimal elevation difference was observed across the site, with the highest and lowest altitude being 211.1 m and 209.4 m AHD respectively. Average annual rainfall for the region is 413 mm, with interpolated. Soil sampling at the site was undertaken in April of 2017

4.2.3. Sampling methods

The 60 m x 60 m sample grid, provided 300 sample locations for the entire site with two intact soil cores extracted to 60 cm at each site; one for chemical analysis and the other for BD and moisture measurements. Cores were extracted using a core sleeve with an internal cutting tip diameter (ID) of 43 mm on a utility-mounted hydraulic coring apparatus with attached jackhammer. Each core was sectioned into key depths of 0–10 cm, 10–20 cm, 20–40cm and 40–60 cm, assumed to define the zone of bulk rooting density (i.e. not extent of rooting) for the crops, and stored in sampling bags. The 1200 samples were measured for soil pH, exchangeable sodium percentage (ESP), BD and cation exchange capacity (CEC), along with other soil structural and chemical measurements not used in this study in accordance to Rayment and Lyons (2011). A summary of the soil properties used for providing the agronomic recommendations is presented in Table 4.2.

Table 4.2. Summary of directly measured soil attributes for all depth increments

	0–10 cm				*10–20 cm*				*20–40 cm*				*40–60 cm*			
	Min	*Max*	*Average*	*SD*	*Min*	*Max*	*Average*	*SD*	*Min*	*Max*	*Average*	*SD*	*Min*	*Max*	*Average*	*SD*
pH	5.27	9.15	6.58	0.64	5.98	9.23	7.52	0.68	6.55	9.45	8.23	0.57	5.98	9.65	8.72	0.56
BD	1.18	1.83	1.47	0.11	1.37	1.84	1.61	0.08	1.01	1.85	1.64	0.08	1.1	1.91	1.68	0.07
CEC	5.78	38.28	16.55	6.39	7.64	39.21	23.31	6.16	10.07	66.5	28.24	5.46	11.05	41.98	29.57	4.53
ESP	0.03	20.86	4.01	3.17	0.13	26.21	5.32	3.75	0.05	30.33	7.36	4.69	0.14	34	10.48	5.9

4.2.4. Proximally sensed environmental covariates

Apparent electrical conductivity (EC_a) measurements were taken in 2015 using an on-the-go DualEM^{TM} sensor at 24 m swathe widths. The DualEM^{TM} sensor provided depth-weighted integrations of EC_a measurements at 4 depth increments of 0–25, 0–75, 0–125, and 0–275 cm. Land elevation was also collected at 24 m swath widths using Real-time Kinematic (RTK) GPS equipment calibrated for <2 cm accuracy. Crop yield data for the 2013-2016 seasons were measured using a harvester-mounted yield monitor and collected at 12 m swathe widths. These were geographically referenced using RTK GPS equipment, with a 2 cm accuracy. OK was used to derive a 6 m spatial interpolation for ECa, elevation and crop yield at the site.

4.2.5. Spatial prediction methods

4.2.5.A.. Random transect sampling

Random transect sampling was simulated to obtain a paddock average of soil conditions representing common agronomic practice for Australian agricultural fields. This average was used to calculate the average homogenous field rate application — colloquially referred to as ‘blanket-rate’ (BR) application. Transects were selected by randomising the start and end member location of the sampling transect along the baselines (Figure 4.1).The baselines were at a distance of 30 m parallel to the north-eastern and south-western paddock boundaries. The *N* samples were subsequently located at equidistant locations as a line between the defined end members to achieve each sampling density. In order to simulate observed conditions at the identified sampling locations OK was undertaken to interpolate the observed dataset to the simulated locations. To not bias the result to a single transect, this process was repeated 10 times for each sampling density such that a range of agronomic recommendations could be obtained for 10 different transect locations, thus providing insight into the sensitivity of

agronomic recommendations to random transect location selection. An example of one transect iteration for sampling density *N*=20 is displayed in Figure 4.1.

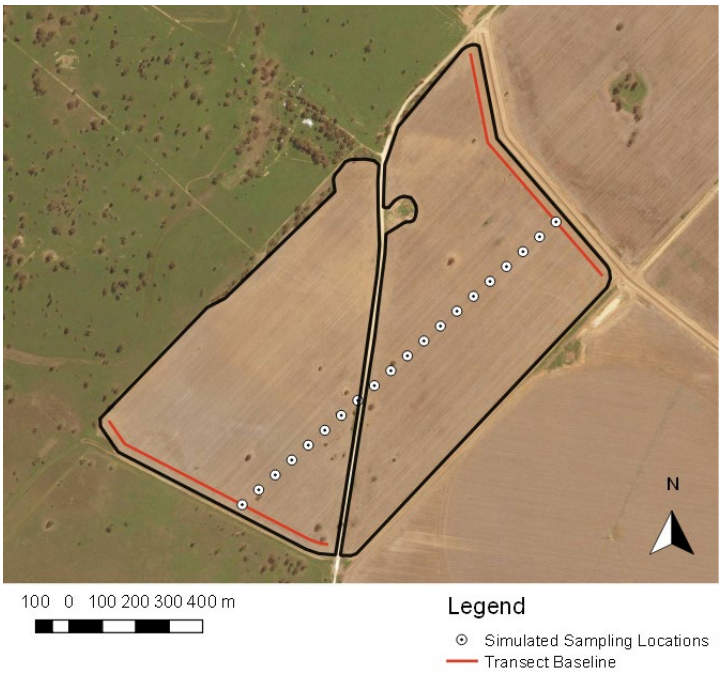


Figure 4.1. Example of 1 simulation of random transect selection with N = 20 samples.

.

4.2.5.B. Management zone sampling

Spatial management zones were identified using a k-means clustering approach. The k-means algorithm identified five management zones (Figure 4.2) using all available auxiliary environmental covariates (i.e. 2013–2016 crop yield, EC_a measurements at 4 depth integrations and elevation). These data layers provided some indication of the level of inherent spatial variability within site, however, provide limited ability to diagnose the cause of the variability. K-means clustering aims to partition datasets into discrete subgroups, such that within-cluster variance is minimised, whilst variance between the centroids or means of each cluster is maximised. Within-cluster variance was minimised by application of the squared error function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \qquad \text{Equation 4.4.}$$

where k is the number of pre-defined clusters, n is the total number of data points in cluster i, x is the observation data point, and c is the centroid of cluster j. From the identified

management zones, a random sampling procedure was simulated to provide zone averages of soil conditions. The number of zones was kept static across all simulations, with sampling density within each zones being *N*/5.

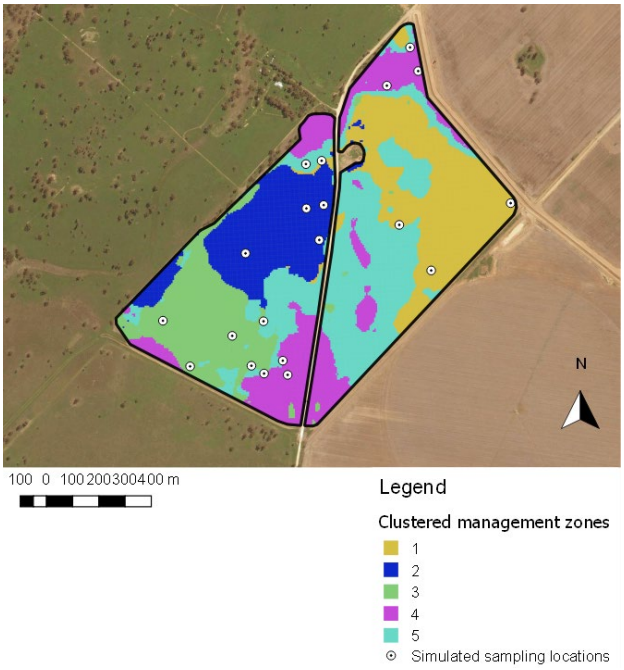


Figure 4.2. Spatial clusters identified using k-means clustering based on 9 environmental covariates. Sampling locations shown for 1 simulation at sampling density N = 20.

4.2.5.C. Ordinary Kriging

OK was used in the spatial prediction of soil properties. A stratified random sampling procedure was employed to identify the sample locations, such that the total site was well represented. For each sampling density *N*, a total of *N* strata were identified using the k-means based *spcosa* package in the R programming environment (Walvoort et al., 2010). Sampling locations were selected randomly within each strata (see Figure 4.3). OK was applied to the sampled data to spatially interpolate to a common 6 x 6 m grid according the following formulation:

$$S^j(x_0) = \sum_{i=1}^{n} \lambda_i S^j(x_i) \qquad \text{Equation 4.5.}$$

where S is soil property j at location x_0, n is the number of observations surrounding x_0 used to predict $S(x_0)$, λ_i are the kriging weights and $S^j(x_i)$ is the measured soil property j at location x_i. Kriging weights were found via minimizing the variance error at the prediction

point by fitting a variogram to the semivariances of the data. The fitted variogram was subsequently applied to perform spatial predictions. This was achieved by application of the *automap* package in the R programming environment (Hiemstra and Hiemstra, 2013). Whilst it has been noted in the literature that variogram estimation for accurate kriging requires a minimum of 100 points (Webster and Oliver, 1992), variograms were fitted at sampling densities as low as 10, to observe the effect poor variogram fitment has on final agronomic recommendations.

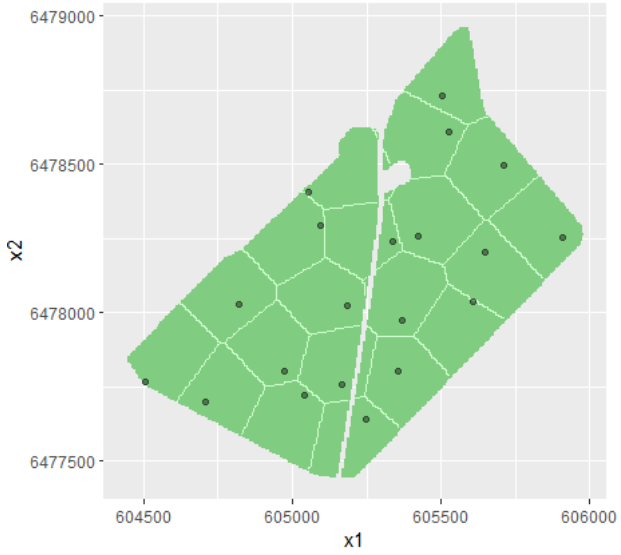


Figure 4.3. Example of site stratification for random selection of N = 20 samples.

4.2.5.D. SSPF regression kriging

For each sampling density *N*, an SSPFe was fitted between the soil properties and the available auxiliary information at *N* locations, with the residuals of the model subsequently being kriged across the geographic space and added to the predictions. The developed SSPFe for any given location *x,y* is given as:

$$S_a = f(EC_a^1, EC_a^2, EC_a^3, EC_a^4, Yield^{2013}, Yield^{2014}, Yield^{2015}, Yield^{2016}, elevation) + e \quad \text{Equation 4.6.}$$

where S_a is the soil attribute, EC_a are the electrical conductivity (EC) readings at the 4 depth integrations, *Yield* represents crop yield measurements, elevation is the elevation above sea level at point *(x, y)* and *e* is the model error.

For each sampling density, *N*, a conditioned Latin-hypercube approach (Minasny and McBratney, 2006b) was adopted to select *N* training examples such that the feature space of the environmental covariates was appropriately represented by maximally stratifying the

environmental covariates (Minasny and McBratney, 2006a) (Figure 4.4). This was achieved by application of the *cLHS* package in the R programming environment (Roudier et al., 2012). A multiple linear regression model was fitted to the mean-normalised environmental covariates for prediction of pH, ESP, BD and CEC. The training model was subsequently employed to predict all soil attributes across the site, with the addition of the kriged residual for each location.

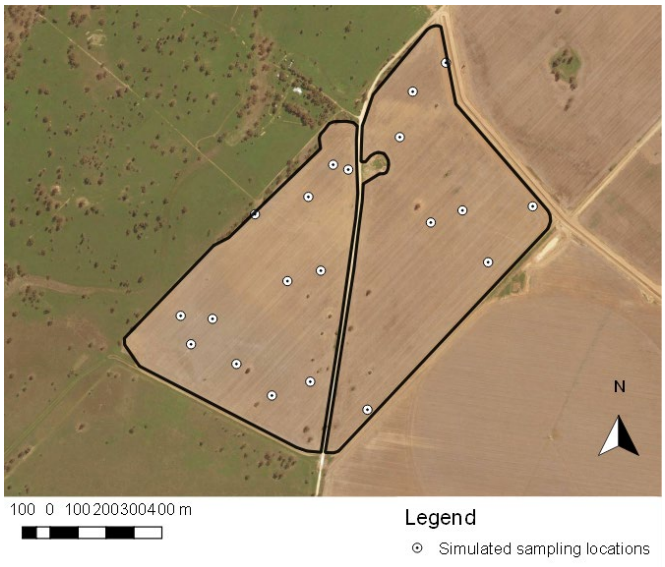


Figure 4.4. Example of sample site selection using cLHS for 1 simulation of N = 20 sampling density.

4.2.6. Gypsum and lime application calculations

The developed DSMs were used to calculate gypsum recommendations for the site to investigate the agronomic and economic consequences of each sampling method under various sampling densities. The widely accepted gypsum recommendation formula devised by Oster and Jayawardane (1998) was adopted and is given as follows:

$$GR = 0.0086.\rho_b.d.CEC.(ESP_i - ESP_j) \quad \text{Equation 4.7.}$$

where ρ_b is the BD in Mg/m^3, d is the depth to be treated in m, CEC is CEC in mmol$_c$/kg, ESP_i and ESP_j are the observed and target soil ESPs. A value of ESP_j=3, as guided by Shainberg et al. (1981), was used to provide a target benchmark for soil dispersion amelioration at all locations, with a calcium exchange efficiency factor of 75% (Bennett et al., 2016). The assumed baseline gypsum requirement — gypsum requirement calculated form the 300 core 60x60 m regular grid —is given in Table 4.3 and Figure 4.5.

Lime recommendation rates were calculated using equation x. A conversion factor of 0.26 for a clay soil and target pH of 7.5 to reduce to neutral were adopted as per Lisa (2019).

$$Pure\ lime\ requirement\ (t/ha) = \frac{Target\ pH - Current\ pH}{Conversion\ factor} \qquad \text{Equation 4.8}$$

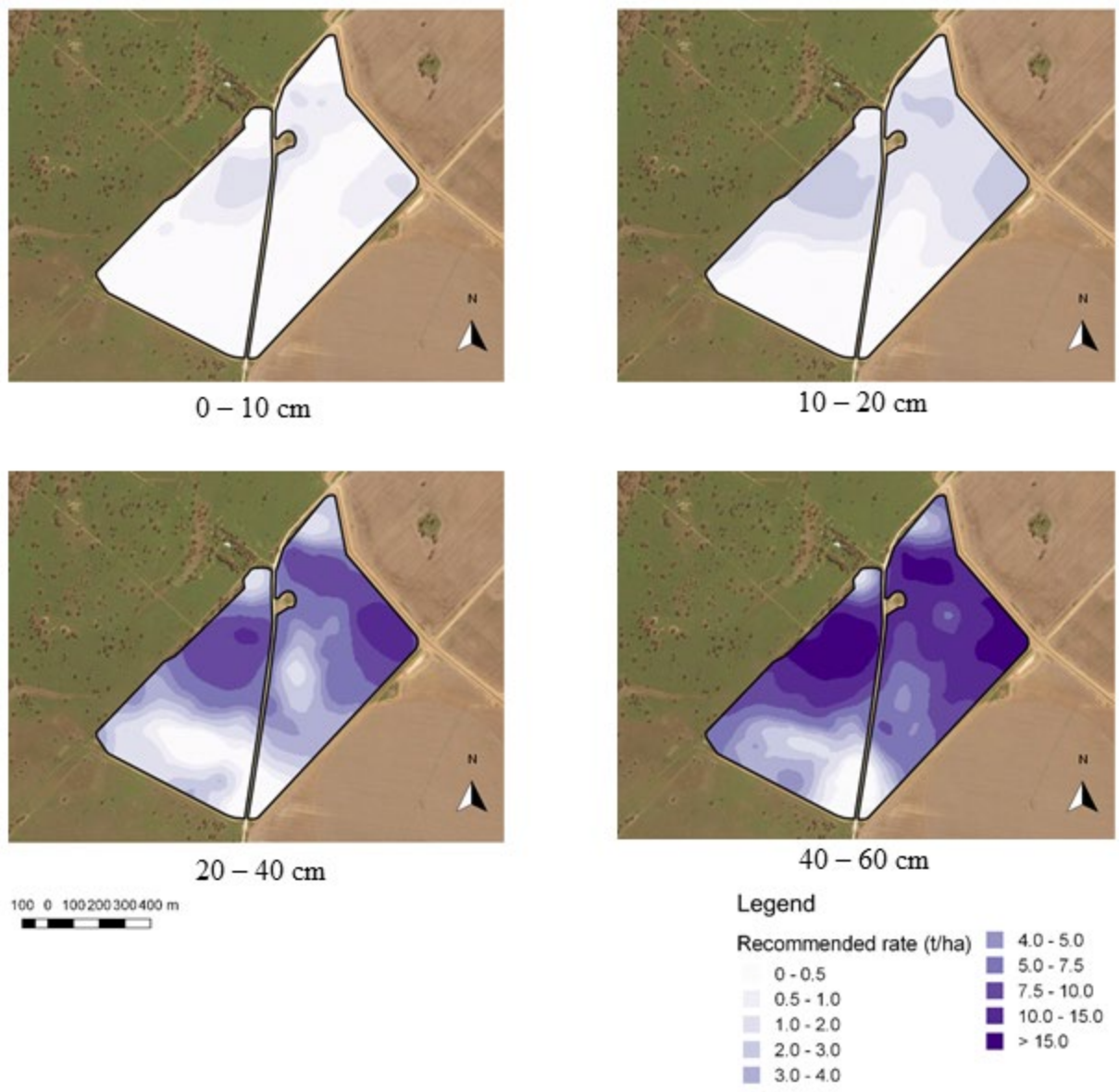


Figure 4.5. Actual spatial gypsum recommendation based on observed samples for the 4 depth increments.

Table 4.3. Summary of gypsum recommendation for the site at all depths

Depth	*Total (t)*
0 – 10 cm	36.5
10 – 20 cm	109.7
20 – 40 cm	517.05
40 – 60 cm	952.85

4.3. Results

4.3.1. Accuracy of spatial prediction methods

The ability to characterise and map soil pH and ESP at the investigation site, of the 4 methods examined, is summarised in Figure 4.6, with multiple sampling densities presented. It is evident for all sampling densities that OK prevails over other methods in terms of characterising pH and ESP at the investigation site, with the exception of the clustering method at sample density 10. Increases in sampling density do not appear to greatly change the spatial prediction errors of the random transect and clustering methods, however, are greatly influential on the predictive accuracy of the kriging and SSPFe methods. For these two methods, the prediction accuracy greatly improves to a density of 50 samples, after which, only minimal improvements are achieved.

The accuracy of spatial predictions are correlated with the degree of spatial variance present in the predicted layer. For soil pH predictions, accuracy generally increases with depth for all methods. This result is expected, as the spatial variance of soil pH also generally decreases with depth, with the exception of the 10–20 cm layer (Table 4.2). For ESP however, spatial variance increases to depth, which explains the decreased prediction accuracy of all methods as depth increases.

Sensitivity of the models to random initialisation was tested and is represented by the range bars presented in Figure 4.6 and Figure 4.7. At low sampling densities (i.e. ≤ 20), the SSPFe and OK methods are most sensitive to sample selection, therefore expressing the largest degree of uncertainty. The uncertainty expressed by the SSPFe at a sampling density of 10 however is substantially greater than other methods, suggesting that the accuracy of regression kriging approaches are highly sensitive to calibration sample selection at low sampling densities. The uncertainty of the transect method is consistent across all sampling densities, and represents the greatest uncertainty at sampling densities ≥ 50.

Correlations between the individual environmental covariates and each predicted soil property are presented in Table 4.4. In general, the environmental covariates do not correlate well with pH and ESP. The correlation between yield and pH and ESP is stronger in the surface layers, whereas the correlation of EC_a measurements are greatest in the subsurface layers.

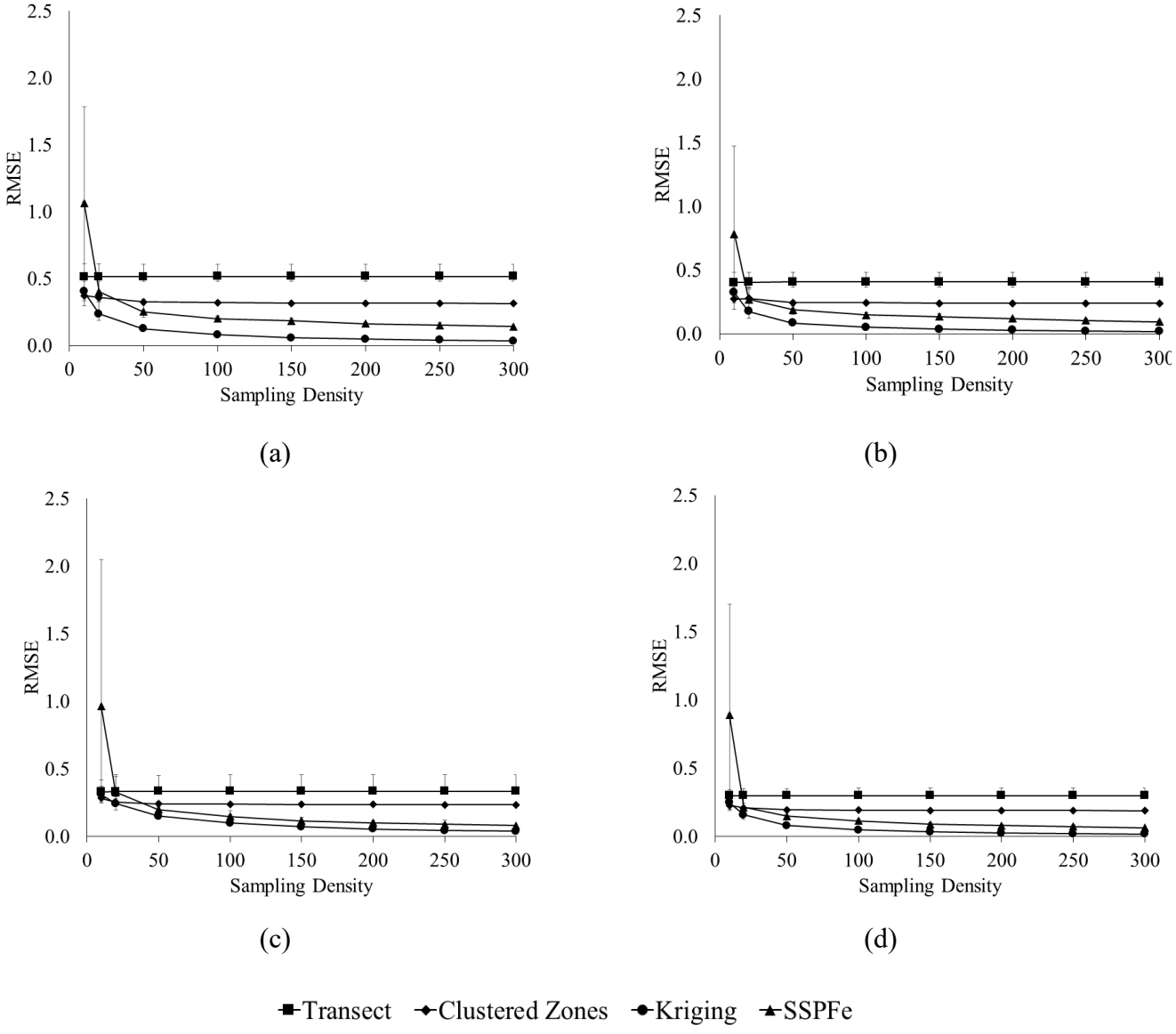


Figure 4.6. Mean RMSE of soil pH predictions over 10 simulations at depths 0–10 cm (a), 10–20 cm (b), 20–40 cm (c) and 40–60 cm (d) for the 4 sampling methods investigated. Bars represent the RMSE range for 10 simulations for each sampling density (x samples /108 ha).

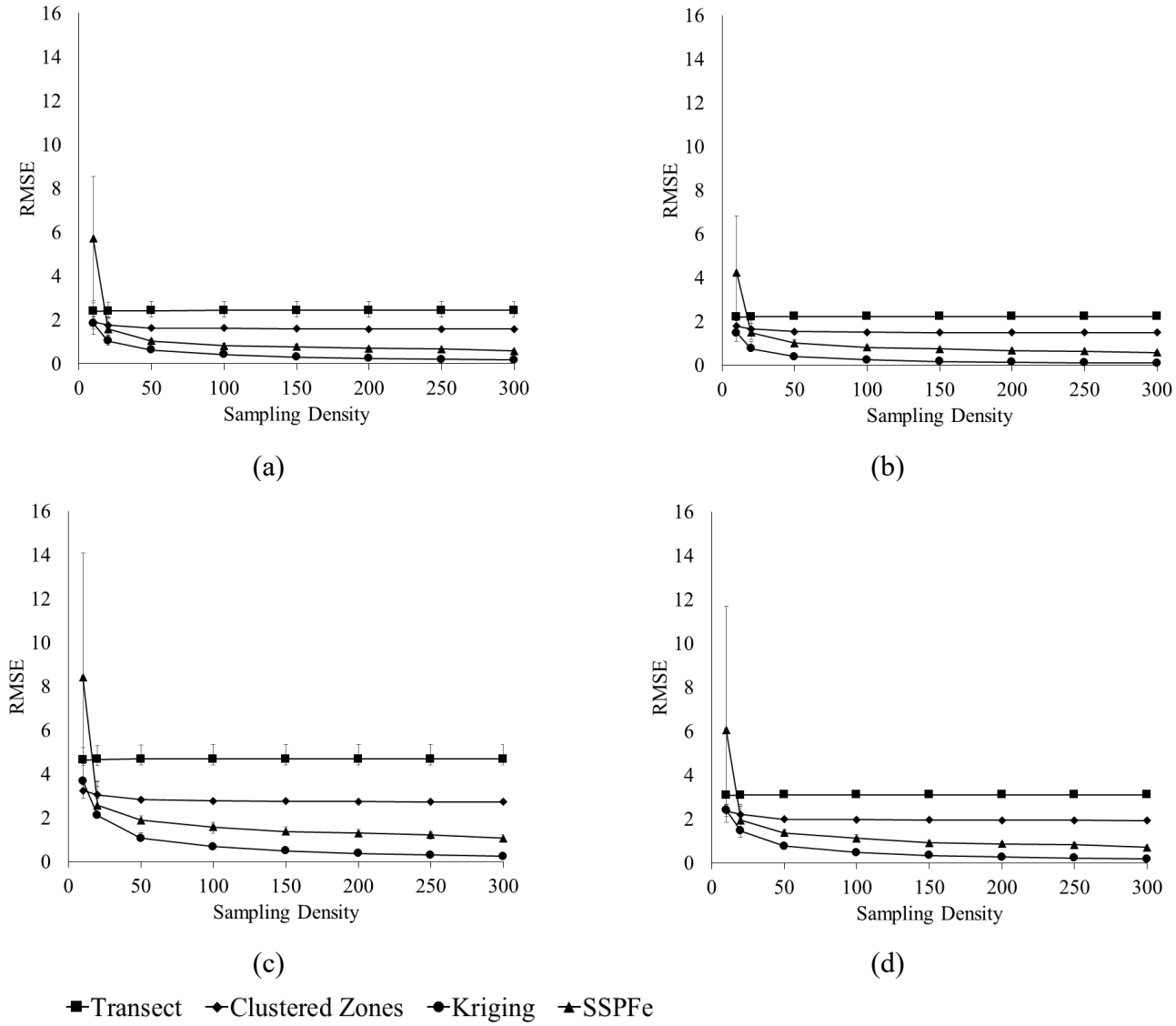


Figure 4.7. Mean RMSE of soil ESP predictions over 10 simulations at depths 0–10cm (a), 10–20 cm (b), 20–40 cm (c) and 40–60 cm (d) for the 4 sampling methods investigated. Bars represent the RMSE range for 10 simulations for each sampling density (x samples /108 ha).

Table 4.4. Correlation coefficients for environmental covariates and soil properties used in the development of the SSPFe. Subscript 1–4 represents depth layers 0–10 cm, 10–20 cm, 20–40 cm and 40–60 cm, respectively. Highlighted cells contain correlation coefficients ≥0.5.

	pH_1	pH_2	pH_3	pH_4	ESP_1	ESP_2	ESP_3	ESP_4
2013 Yield	0.47	0.48	0.37	0.27	0.20	0.27	0.23	0.33
2014 Yield	0.41	0.25	0.13	0.03	0.05	0.02	0.04	0.08
2015 Yield	0.32	0.26	0.18	0.01	0.18	0.13	0.13	0.17
2016 Yield	0.39	0.22	0.06	0.16	0.16	0.14	0.14	0.19
0–25 cm EC_a	0.07	0.36	0.49	0.49	0.58	0.63	0.68	0.78
0–75 cm EC_a	0.04	0.33	0.49	0.51	0.57	0.63	0.68	0.78
0–125 cm EC_a	0.07	0.35	0.50	0.52	0.55	0.61	0.66	0.77
0–275 cm EC_a	0.03	0.31	0.49	0.52	0.52	0.59	0.64	0.74
Elevation	0.52	0.27	0.08	0.24	0.44	0.36	0.40	0.41

4.3.2. Spatial prediction errors

Prediction error maps for soil pH and ESP are displayed in Figure 4.8 and Figure 4.9, respectively. The spatial errors are of a higher magnitude at low sampling densities in the surface layer for pH, and the 40–60 cm layer for ESP. This agrees with the RMSE results of each prediction. The errors for both pH and ESP are spatially correlated, meaning that the areas of large error are generally spatially consistent across sampling methods and densities for a given depth. For example, this is seen in the 0–10 cm layer for soil pH, where all methods severely under predicted values in a region to the south-west of the site, which correlates with zone 3 identified by k-means clustering. The spatial distribution of errors for the SSPFe method, however, did not always agree with the other methods. This is seen in the 40–60 cm depth layer for ESP. Here the higher magnitude errors are less spatially correlated, and occur in smaller, more spatially irregular pockets in comparison to other methods.

Increases in sampling density did not greatly affect the spatial distribution of errors for the random transect and clustered zone methods for soil pH or ESP. This agrees with the RMSE findings in Figure 4.6 and Figure 4.7. For the OK and the SPPFe methods however, the magnitude of error is considerably less and closer to the prediction as sampling density increased from 10 to 50 samples. These errors also become more spatially distributed. The error maps for all sampling densities tested are displayed in the Appendix.

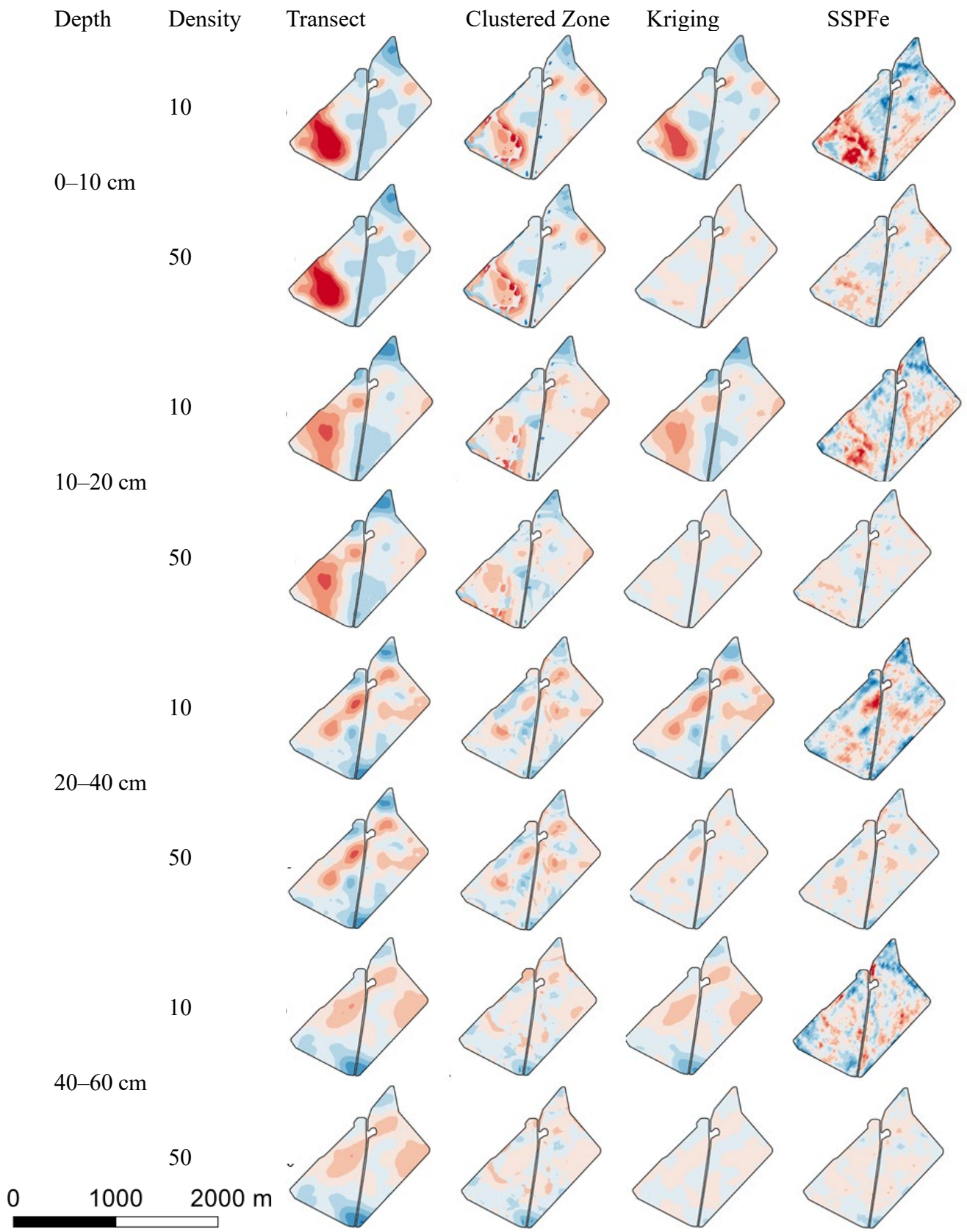


Figure 4.8. Mean prediction error maps of the 4 methods investigated for soil pH at to 60 cm. Error maps shown for sampling densities N = 10 and 50 . Red shades represent under prediction whilst blue shades represent over prediction.

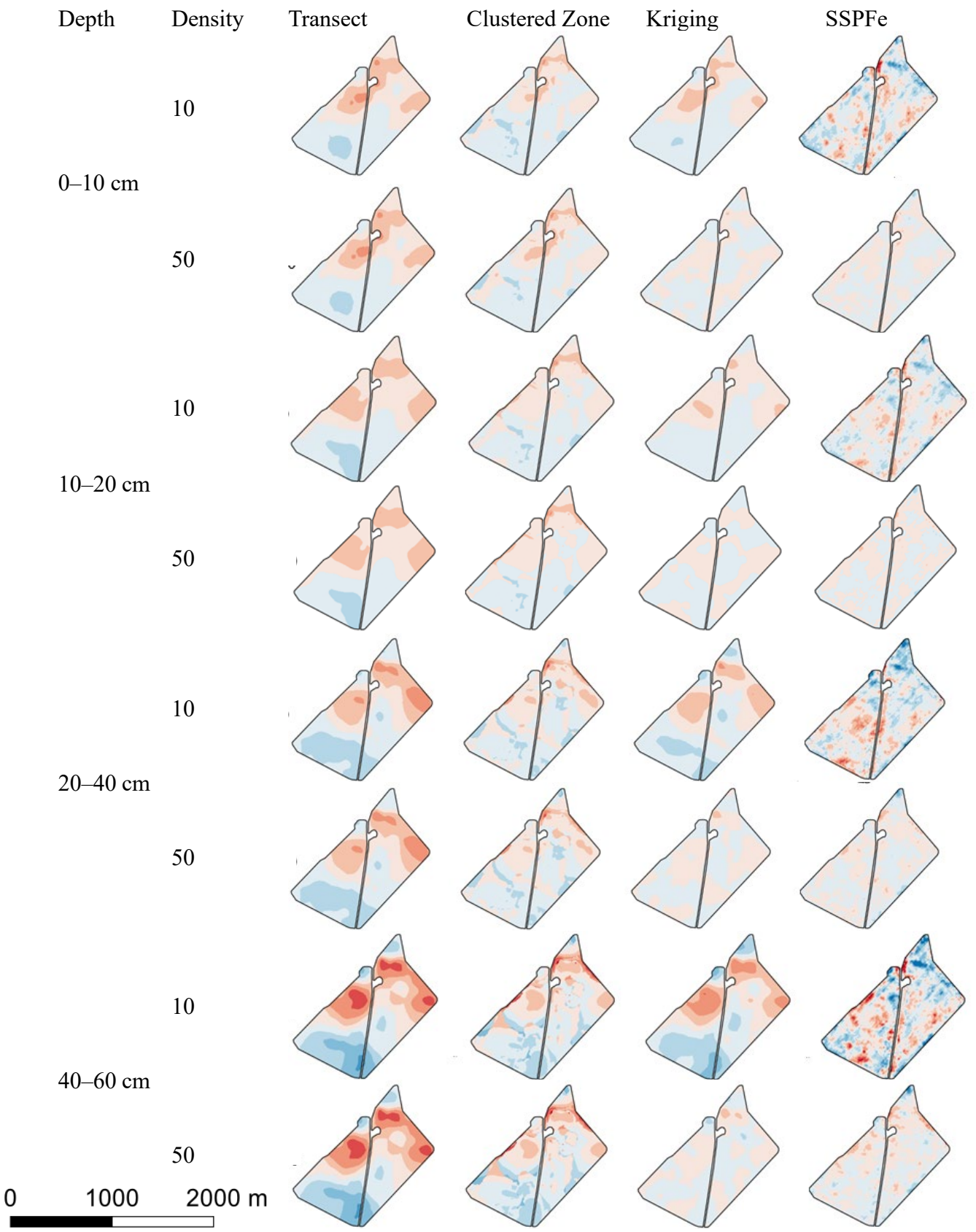


Figure 4.9. Mean prediction error maps of the 4 methods investigated for soil ESP at to 60 cm. Error maps shown for sampling densities N = 10 and 50. Red shades represent under prediction whilst blue shades represent over prediction.

4.3.3. Error of agronomic recommendations

For each simulation of the 4 spatial prediction methods, gypsum and lime recommendations were calculated against the spatial resolution to observe the agronomic consequences of the prediction errors on individual soil properties (Figure 4.10 and Figure 4.11). These agronomic errors were estimated by calculating the net over- and under-application of the amendment for the 8 sampling densities. The net error trend of gypsum and lime recommendations generally reflected that of the RMSE calculations, with error reducing as sampling density increased.

Error trends of gypsum and lime recommendations, based on the spatial predictions, generally agreed with that of the RMSE findings (Figure 4.10 and Figure 4.11). At sampling densities ≥20, the OK method had lower magnitude errors in under and over application of amendments. The magnitude of application error for the OK and SSPFe methods decreased with increasing sampling densities, with errors remaining relatively consistent for the clustered zone and random transect methods. Interestingly, small changes in RMSE of ESP and pH predictions translated into large recommendation errors, in terms of both under- and over-application. This suggests that the accuracy of recommendations are highly sensitive to small changes in spatial prediction performance.

The magnitude of under-application error generally exceeded that of over application for all methods, suggesting that the models under predicted ESP and pH values. Recommendations produced using the bulked transect sampling method were the most inaccurate for both lime and gypsum, with OK producing the best results. The SSPFe method produced highly inaccurate recommendations at the sampling density 10, however this error quickly decreased within increasing sampling density. For both the SSPFe and OK soil prediction methods, the errors greatly improved to a sampling density of 50 samples, after which minimal improvements were observed.

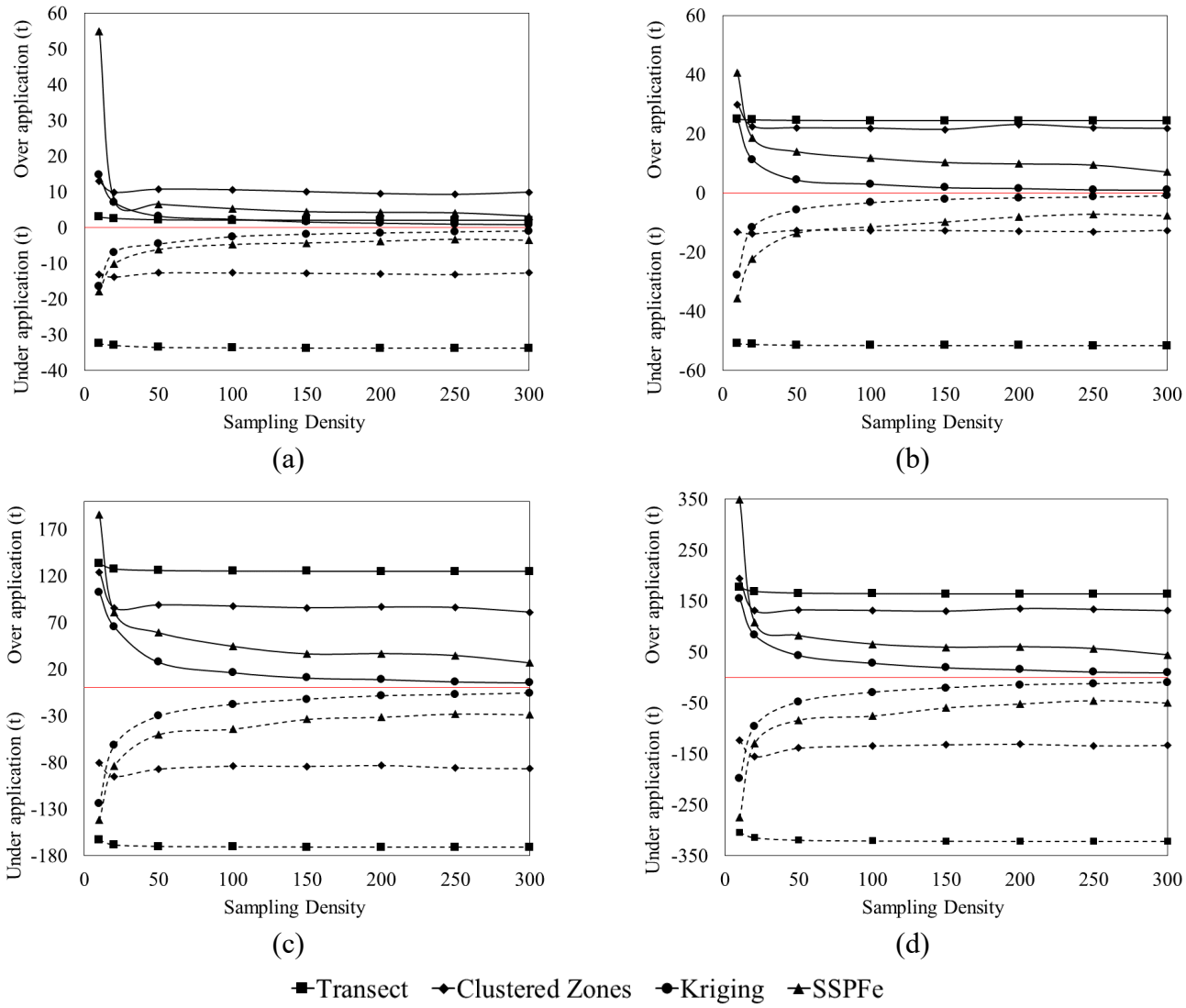


Figure 4.10. Summary of gypsum application recommendations of 4 depth increments of 0–10 cm (a), 10–20 cm (b), 20–40 cm (c) and 40–60 cm (d), based on the spatial predictions of the 4 methods investigated over various sampling densities, in tones (t) of product. Solid and dashed lines represent the over and under application of gypsum for the site respectively (x samples/108 ha).

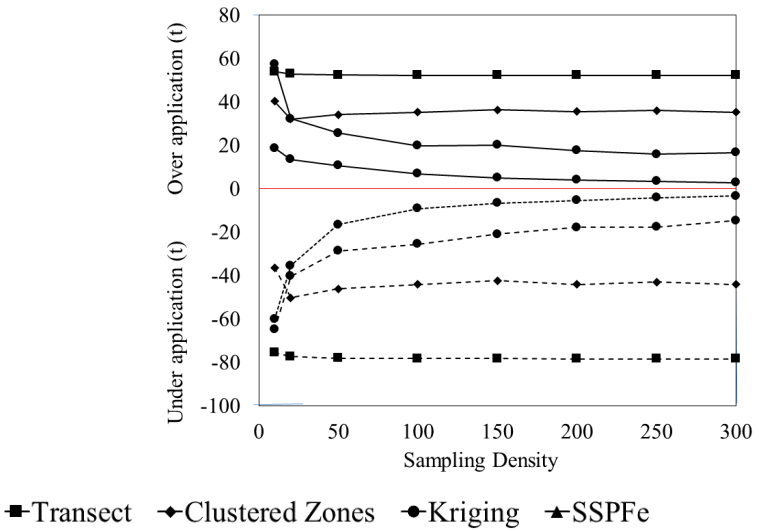


Figure 4.11. Summary of lime application recommendations of the 0–10 cm surface layer based on the spatial predictions of the 4 methods investigated over various sampling densities. Solid and dashed lines represent the over and under application of gypsum for the site respectively (x samples/108 ha).

4.4. Discussion

4.4.1. Agronomic consequences of data limited recommendations

The bulked transect sampling method was used in this study to represent an industry agronomic standard practice. However, it is worth noting that this level of data collection likely over estimates the level of sampling commonly undertaken for agronomic decision making (Lobry de Bruyn and Andrews, 2016) and Bennett and Cattle (2013) suggest this actually over estimates the level of sampling undertaken for agronomic decision making. The results conclusively established that bulked transect sampling was highly inaccurate in representing the site variability and subsequent site-specific gypsum and lime recommendations were highly inaccurate. By extension, this suggests that the current industrial agricultural sampling strategies, which are likely more conservative than bulked transect sampling, would result in significant error pertaining to the resultant recommendations. Gypsum and lime recommendation errors were much greater for bulked transect sampling than that of other methods, with over application magnitudes reaching almost 200 t and 300 t respectively in the 40–60 cm layer for the entire site. With gypsum application costs of approximately $110/t (transported and spread; Bennett et al. (2015a), this error presents great economic significance.

Of potentially greater concern than the cost of over-application is the failure to spatially address the ESP and pH constraints, which would impact on yield potential. Much of the site

was recommended an insufficient ameliorant quantity using the bulked transect method, with under-application being in the order of 25–10 t for gypsum and lime in the 0–10 cm surface layer, and 100–300 t in the 40–60 cm subsurface layer. This error is concerning, considering the level of investment that would be committed on the basis of these recommendations. The long-term lost yield opportunity of this shortfall is likely an important consideration within a long-term amelioration strategy as the insufficient application may not actually result in yield increase, and the site-specific yield potential at the site cannot be realised. Therefore, using a bulked transect sampling approach for site-average recommendations is highly detrimental to the long term agronomic and economic performance of a farming unit, and should be avoided when providing recommendations.

Utilising a spatial sampling strategy to allow for zone management offers improved recommendations over the bulked transect method, with over and under application of gypsum being reduced by approximately 20% and 60% respectively over all sampling densities. This improvement is achieved by reducing the within-zone variance in an attempt to utilize spatial auxiliary information to identify 'homogenous' zones within a site that exhibit similar soil characteristics (Ruß and Kruse, 2011). In doing so, it is assumed that soil properties and yield imitating factors are consistent within each identified zone (Doerge, 1999), which is an incorrect assumption. Whilst the error of this assumption is less than that of spatially averaging soil properties across an entire site, its magnitude is of great agronomic and economic significance, with over- and under-application of gypsum approaching 130 t in the 40–60 cm subsurface layer for the investigated site. This is the result of fitting hard boundaries to continuous soil properties in an attempt to simplify the representation of soil variance. Whilst this allows for improved recommendations with a minimal increase in the data investment, the error of these recommendations remain large in comparison to other methods.

Management zone delineation for variable rate recommendation is further limited by the assumption that all soil properties are spatially correlated with each other and can be represented by the same set of global boundaries. In reality, soil properties may not only vary independently of each other, but may also vary independently to depth, meaning that global boundaries cannot accurately characterize soil characteristics across multiple properties. This behavior was observed at the investigation site (Figure 4.12), where surface pH was varied independently of subsurface pH and both surface and subsurface ESP. Whilst some spatial correlation can be observed between subsurface pH and ESP (R^2=0.74 exponential correlation;

results not shown), these are not well represented by the identified management zones. This presents a second limitation of clustering zone management, which assumes all soil properties are spatially correlated with the auxiliary information used to delineate the zones (e.g. yield data, elevation data and EC data). Zone management therefore over simplifies the detection of spatial variability by providing limited consideration towards the independent variability soil properties and level of spatial correlation that exists at the site.

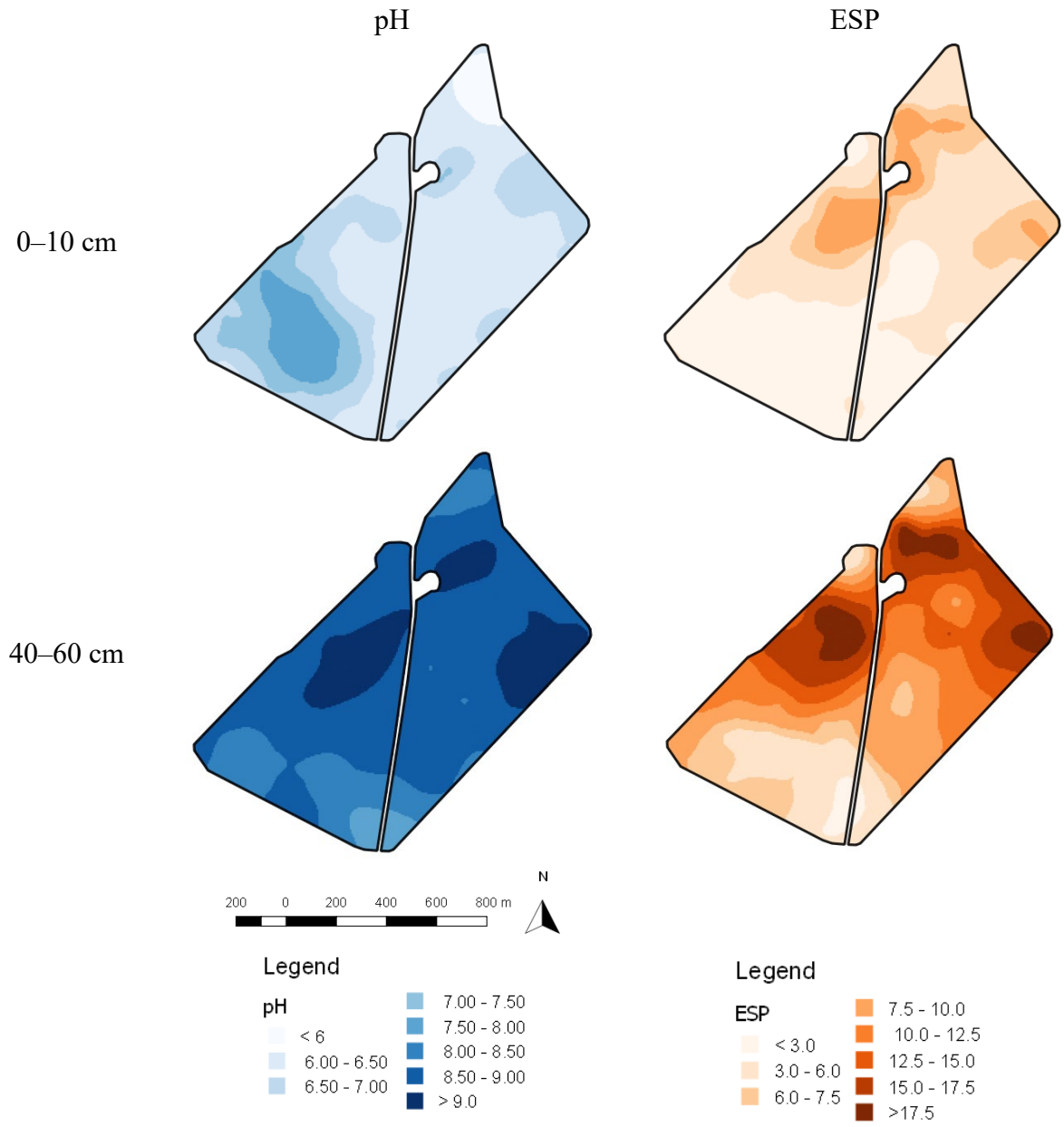


Figure 4.12 Measured soil pH and ESP maps for the investigation site within the 0–10 cm and 40–60 cm depth layers.

Zone management is the current accepted standard for variable rate recommendations for commercial precision agriculture (Li et al., 2007; McBratney et al., 2005; Robertson et al., 2012; Whelan and McBratney, 2003). However it is shown here to result in severe under- and over-application of soil amendment. The economic and agronomic effects of this are significant, due to both wasted resource, and failure to accurately address soil condition to overcome constraints. Agricultural technology is currently capable of applying soil amendment at a much finer scale than what is currently practiced in zone management, with some machines offering row-specific control (<1 m) (see John Deere ® RowCommand^{TM} – www.deere.com) with the aid of sub 2 cm accurate RTK (real-time kinematic) GPS technology. Therefore, whilst zone management offers improvements of BR applications, it's data minimalist approach is currently limiting precision agriculture by failing to identify variation in soil condition at the scale in which it can be managed using more advanced approaches.

4.4.2. Improving recommendations through advanced spatial prediction methods with increased sampling requirements

Spatial variable rate recommendations can be improved significantly by adopting DSM methods that predict soil properties as a continuous function at fine spatial scales across a paddock. This was achieved for both the OK geostatistical method and SSPFe method for the investigation site. These DSM methods offer improvements over zone-based predictions by removing hard separating boundaries between changes in soil condition and expressing these as a continuous function (McBratney et al., 2003). Furthermore, the spatial variation of each soil property allows for the independent treatment of variables, thus overcoming the assumption of spatial correlation between properties and depths layers. However, these improvements are not consistent at low sampling densities (i.e. ≤10), with the zonal management using hard boundaries periodically offering improvements in under prediction errors; this indicates the unreliability of continuous functions at low sampling densities, which really should be expected (Abbaspour et al., 1998; Ahrens, 2008; Burgess and Webster, 1980a). Whilst zonal management periodically provides the best method for spatial predictions at very low sampling densities (i.e. ≤10), representing a low-cost solution, this does not equate to it being the optimal solution in terms of the return on investment of the sampling and amelioration. Prediction and recommendation errors greatly increase with increasing sampling density and surpass that of bulked transect and zone sampling above 20 samples, or approximately 1 sample per 5 ha generic density.

The DSM methods were highly sensitive to increases in sampling density, with the majority of prediction improvement being achieved at 50 samples per 108 ha, from which minimal improvement (error reduction) is made thereafter. This suggests that the spatial variability at the site can be accurately characterised using a sampling density of 1 sample per 2 ha, and is practically meaningful at 1 sample per 5 ha. This density however does not guarantee an economically optimized site characterisation, as the cost of data acquisition is not considered against the economic benefit in terms of both the sampling cost and expected yield return (see Chapter 6 for this discussion).

At low sampling densities (i.e. ≤10), the SSPFe method is highly inaccurate, suggesting that 10 calibration samples is not sufficient to obtain a site specific relationship between the environmental covariates and ESP or pH. At this sampling density however, it is not expected to achieve an appropriately fitted model to the data due to the inherent complexities within the soil system. In fact, achieving acceptable model calibration may not be expected at increased sampling densities of 20 or 50, with SSPFe models often being calibrated at much larger densities (Cockx et al., 2010; Florinsky et al., 2002; Li, 2010; Malone et al., 2018; Niang et al., 2014; Pantazi et al., 2016). Furthermore, OK is rarely applied at these low sampling densities due to the inaccuracies of fitting a variogram model (Bishop and McBratney, 2001), with Webster and Oliver (1992) reporting that 100 samples are the absolute minimum required for an appropriate variogram fit. Whilst a data density of 20 or 50 samples may not be considered sufficient for SSPFe development or variogram fitting from a pedometric perspective, the agronomic error or these methods remains less than that of bulked transect and zone management methods at an equivalent density. Therefore, the context of the spatial prediction problem must be considered when determining acceptable sampling densities.

At all sampling densities, OK produced superior results over the SSPFe method, which is in direct contrast to that found by Odeh et al. (1994), Odeh et al. (1995), Hengl et al. (2004) and Bishop and McBratney (2001). SSPFe methods rely on the spatial correlation between soil properties and environmental covariates used in predictions. However, for the site investigated, little correlation exists between the available environmental covariates and pH or ESP. The RMSE of predictions were similar between the SSPFe and OK methods where a greater correlation existed between the response variable and the environmental covariates (e.g. pH and ESP in the 40–60 cm layer). The lack of correlation between ESP, or pH, and the environmental covariates suggest that other soil properties have greater influence on their

values, and cannot be used to describe the variability of ESP or pH. Concomitantly, it may be that a greater array of environmental covariates could have improved the relationship with pH and/or ESP. Therefore, SSPFe methods only offer improved performance in situations where environmental correlation is present, which is dependent on a site's inherent characteristics (e.g. management history) and constraints, as well as the available environmental covariates (e.g. NDVI imagery, gamma-radiometrics, elevation, yield, ECa etc.).

The spatial prediction accuracy of SSPFe can also be improved by employing more sophisticated non-linear machine learning (ML) techniques that are capable of detecting complex relationships between soil properties and environmental variables. These have been applied in the literature using artificial neural networks (ANNs) (Florinsky et al., 2002; McKenzie and Ryan, 1999; Minasny and McBratney, 2010), regression-trees (Henderson et al., 2005; Lacoste et al., 2014; McKenzie and Ryan, 1999), SVMs (Ballabio, 2009; Were et al., 2015) and ANNs (Behrens et al., 2005; Chang and Islam, 2000; Dai et al., 2011). However, the data requirements of these methods are exponentially increased over linear methods, due to the absence of structural assumptions within the data that subsequently allows for higher complexity. In this study, these non-linear methods are not suitable for SSPFe development given the data volume in comparison to the complexity of the problem (a data limited environment), and may only offer improvements when the size of calibration sampling is large.

4.4.3. The effect of sample selection on prediction uncertainty

Each of the four methods investigated employs a random initialization of search parameters that identify the selection of samples used in the spatial predictions. A level of uncertainty therefore exists for each method. Uncertainty generally decreases with increased sampling density, as a greater percentage of the total population is accounted for. The uncertainty of the SSPFe method is the most sensitive of the methods to random initialization of calibration samples at reduced sampling densities (i.e. <50). Whilst the cLHS technique employed to select calibration samples ensures appropriate distribution within the feature space, it cannot guarantee that the selected samples are representative of the relationships that exist between the environmental covariates and soil properties. Model calibration is therefore highly biased towards these samples. This bias can only be reduced by increasing sampling requirements. The magnitude of this uncertainty suggests that SSPFe methods should not be attempted at sampling densities <20, although this is dependent on the inherent variability that exists within the site and the ability for the environmental covariates at the site.

The assumed industry standard method of using a bulked transect was shown to be highly sensitive to the selection of the transect from which samples are taken. Using this method, gypsum recommendations were up to 482% incorrect in the surface layer and 32% incorrect in the 40–60 cm subsurface layer. In practice, the transect is rarely truly randomised, and is instead often selected to simply span the diagonal length of the field, with limited attempt to appropriately represent the inherent variability of the site. This presents a large agronomic concern, as the soil amendment advice that is provided is largely influenced by how the transect was selected. Furthermore, while zonal management provides improvement on this, it has been shown here that by using of the same number of samples within the field, a much improved outcomes where more advanced techniques are used. This is a significant opportunity for agriculture, as it requires no further expense in sampling, but delivers improved predictions with high certainty.

This work has presented a detailed investigation for a single site, and we acknowledge that there are limitations related to this. Specifically, the results obtain should not be expected to directly transfer to new sites. However, this work provides a valuable discussion of the considerations for data density, and should be used over a range of new sites in order to confirm the practically useful recommendation of 1 sample in every 5 ha, as well as the accurate recommendation of 1 sample in every 2 ha, for the spatially continuous SSPFe and cLHS methods.

4.5. Conclusion

To drive on-farm profitability with consideration toward social responsibility of management, the influence of spatial soil variability on the accuracy of VR soil amelioration advice is required to be better understood. Furthermore, it must be understood how this advice is influence by different sampling techniques employed at various levels of soil sampling investment. The results in this study have shown the agronomic advice based on the widely adopted, bulk transect sampling method for blanket rate application of soil amendment is largely inaccurate, leading to potential large and under applications of amendments at significant cost to the grower (either as yield penalty or unnecessary application).. Using this approach, application of gypsum and lime is frequently misapplied to the spatial areas in which it is most required, with total over and under application tonnages within a single field being substantial (i.e >320 t and 560 t respectively for a 108 ha field for treatment to 60 cm). Transect sampling for BR application should be avoided for soil amelioration recommendations

The most accurate applications of lime and gypsum were achieved using a VR approach, based on an OK spatial prediction method. In general, this was achieved over all sampling densities. Whilst traditionally it was been recommended that regression kriging approaches, such as SSPFe, are preferred for spatial predictions due to improved performance (Bishop and McBratney, 2001; Hengl et al., 2004; Odeh et al., 1994; Odeh et al., 1995), the site investigated exhibited minimal spatial correlation between environmental covariates and the individual soil properties pertaining to sodicity and acidity constraint metrics. Therefore, environmental correlation between the auxiliary information and predicted soil attribute must be considered when making spatial predictions. Hence, where this correlation is poor, or at worst, unknown, both OK and SSPFe methods should be investigated simultaneously in providing spatial agronomic advice, for example, using probability sampling validation (Brus et al., 2011)

Sampling density was shown to be highly influential on the recommendation advice for OK and SSPFe methods, however did not contribute to large changes in the error for transect sampling or zone management. Spatial prediction and recommendation accuracies greatly improved to a sampling density of 50 for OK and SSPFe, with minor improvements being achieved thereafter. Selecting the most optimal sampling density requires further consideration towards the economics of increased data collection and its effects on crop performance due to improved spatial agronomic advice for soil amelioration. This will be investigated in Chapter 6 of this thesis.

4.6. References

Abbaspour, K., Schulin, R., van Genuchten, M.T., Schläppi, E., 1998. An alternative to cokriging for situations with small sample sizes. Mathematical geology 30(3), 259-274.

Ahrens, R.J., 2008. Digital soil mapping with limited data. Springer Science & Business Media.

Ballabio, C., 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151(3-4), 338-350.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Journal of plant nutrition and soil science 168(1), 21-33.

Bennett, J.M., Submitted. Soil Security for Australia. Soil Systems.

Bennett, J.M., Cattle, S., 2013. Adoption of soil health improvement strategies by Australian farmers: I. Attitudes, management and extension implications. The Journal of Agricultural Education and Extension 19(4), 407-426.

Bennett, J.M., Cattle, S., 2014. Adoption of soil health improvement strategies by Australian farmers: II. Impediments and incentives. The Journal of Agricultural Education and Extension 20(1), 107-131.

Bennett, J.M., Cattle, S., Singh, B., 2015. The efficacy of lime, gypsum and their combination to ameliorate sodicity in irrigated cropping soils in the Lachlan Valley of New South Wales. Arid Land Research and Management 29(1), 17-40.

Bennett, J.M., Marchuk, A., Raine, S., Dalzell, S., Macfarlane, D., 2016. Managing land application of coal seam water: A field study of land amendment irrigation using saline-sodic and alkaline water on a Red Vertisol. Journal of environmental management 184, 178-185.

Bishop, T., McBratney, A., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103(1-2), 149-160.

Boydell, B., McBratney, A., 2002. Identifying potential within-field management zones from cotton-yield estimates. Precision agriculture 3(1), 9-23.

Brus, D., Kempen, B., Heuvelink, G., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62(3), 394-407.

Brus, D.J., Heuvelink, G.B., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138(1-2), 86-95.

Burgess, T., Webster, R., 1980. Optimal interpolation and isarithmic mapping of soil properties: I. The semivariogram and punctual kriging. Journal of soil science 31(2), 315-331.

Chang, D.-H., Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural network. Remote Sensing of Environment 74(3), 534-544.

Cockx, L., Van Meirvenne, M., Vancoillie, F., Verbeke, L., Simpson, D., Saey, T., 2010. A Neural Network Approach to Topsoil Clay Prediction Using an EMI-Based Soil Sensor, Proximal Soil Sensing. Springer, pp. 245-254.

Dai, X., Huo, Z., Wang, H., 2011. Simulation for response of crop yield to soil moisture and salinity with artificial neural network. Field Crops Research 121(3), 441-449.

De Gruijter, J., Brus, D.J., Bierkens, M.F., Knotters, M., 2006. Sampling for natural resource monitoring. Springer Science & Business Media.

De Gruijter, J., McBratney, A., 1988. A modified fuzzy k-means method for predictive classification.

De Gruijter, J.J., 1977. Numerical classification of soils and its application in survey. Pudoc.

Doerge, T., 1999. Defining management zones for precision farming. Crop Insights 8(21), 1-5.

Florinsky, I.V., Eilers, R.G., Manning, G., Fuller, L., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling & Software 17(3), 295-311.

Fu, Q., Wang, Z., Jiang, Q., 2010. Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO. Mathematical and computer modelling 51(11-12), 1299-1305.

Heath, R., 2018. Editorial to John Ralph Essay Competition 2018: Should society determine the right to farm? Farm Policy Journal 15(5), 2-3.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124(3), 383-398.

Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120(1-2), 75-93.

Hengl, T., Rossiter, D.G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Soil Research 41(8), 1403-1422.

Heuvelink, G.B., Brus, D.J., de Gruijter, J.J., 2006. Optimization of sample configurations for digital mapping of soil properties with universal kriging. Developments in soil science 31, 137-151.

Hiemstra, P., Hiemstra, M.P., 2013. Package 'automap'. compare 105, 10.

Hudson, G., Wackernagel, H., 1994. Mapping temperature using kriging with external drift: theory and an example from Scotland. International journal of Climatology 14(1), 77-91.

Jenny, H., 1941. Factors ofSoil Formation, A System of Quantitative Pedology. McGraw-Hill.

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

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296-311.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water resources research 31(2), 387-398.

Li, Y., 2010. Can the spatial prediction of soil organic matter contents at various sampling scales be improved by using regression kriging with auxiliary information? Geoderma 159(1-2), 63-75.

Li, Y., Shi, Z., Li, F., Li, H.-Y., 2007. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. Computers and Electronics in Agriculture 56(2), 174-186.

Lobry de Bruyn, L., Andrews, S., 2016. Are Australian and United States farmers using soil information for soil health management? Sustainability 8(4), 304.

Lush, D., 2018. Should society determine the right to farm? Farm Policy Journal 15(4), 4.

Malone, B.P., Odgers, N.P., Stockmann, U., Minasny, B., McBratney, A.B., 2018. Digital mapping of soil classes and continuous soil properties. Pedometrics, 373-413.

McBratney, A., de Gruijter, J., 1992. A continuum approach to soil classification by modified fuzzy k ‐ means with extragrades. Journal of Soil Science 43(1), 159-175.

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

McBratney, A., Webster, R., Burgess, T., 1981. The design of optimal sampling schemes for local estimation and mapping of of regionalized variables—I: Theory and method. Computers & Geosciences 7(4), 331-334.

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

McBratney, A.B., Odeh, I.O., 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77(2-4), 85-113.

McBratney, A.B., Odeh, I.O., Bishop, T.F., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97(3-4), 293-327.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89(1-2), 67-94.

Minasny, B., McBratney, A., 2006a. Latin hypercube sampling as a tool for digital soil mapping. Developments in soil science 31, 153-606.

Minasny, B., McBratney, A., 2010. Conditioned Latin hypercube sampling for calibrating soil sensor data to soil properties, Proximal soil sensing. Springer, pp. 111-119.

Minasny, B., McBratney, A.B., 2006b. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & geosciences 32(9), 1378-1388.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301-311.

Müller, W.G., 2001. Collecting Spatial Data: Optimum Design of Experiments for Random Fields, 2nd Ed.

Niang, M.A., Nolin, M.C., Jégo, G., Perron, I., 2014. Digital Mapping of soil texture using RADARSAT-2 polarimetric synthetic aperture radar data. Soil Science Society of America Journal 78(2), 673-684.

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

Odeh, I.O., McBratney, A., Chittleborough, D., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67(3-4), 215-226.

Odeh, I.O., McBratney, A.B., 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. Geoderma 97(3-4), 237-254.

Oster, J., Jayawardane, N., 1998. Agricultural management of sodic soils.

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

Rayment, G.E., Lyons, D.J., 2011. Soil chemical methods: Australasia, 3. CSIRO publishing.

Robertson, M., Llewellyn, R., Mandel, R., Lawes, R., Bramley, R., Swift, L., Metz, N., O'Callaghan, C., 2012. Adoption of variable rate fertiliser application in the Australian grains industry: status, issues and prospects. Precision Agriculture 13(2), 181-199.

Roudier, P., Beaudette, D., Hewitt, A., 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. Digital soil assessments and beyond, 227-231.

Ruß, G., Kruse, R., 2011. Exploratory hierarchical clustering for management zone delineation in precision agriculture, Industrial Conference on Data Mining. Springer, pp. 161-173.

Shainberg, I., Rhoades, J., Prather, R., 1981. Effect of Low Electrolyte Concentration on Clay Dispersion and Hydraulic Conductivity of a Sodic Soil 1. Soil Science Society of America Journal 45(2), 273-277.

Sibson, R., 1981. A brief description of natural neighbour interpolation. Interpreting multivariate data.

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

Vašát, R., Heuvelink, G., Borůvka, L., 2010. Sampling design optimization for multivariate soil mapping. Geoderma 155(3-4), 147-153.

Walvoort, D.J., Brus, D., De Gruijter, J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Computers & Geosciences 36(10), 1261-1267.

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

Wendelberger, J.G., 1981. The Computation of Laplacian Smoothing Splines with Examples, WISCONSIN UNIV-MADISON DEPT OF STATISTICS.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52, 394-403.

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

5. Crop yield prediction using machine learning for the purpose of soil constraint diagnosis

5.1. Introduction

The success of technology to accurately identify the effect of soil constraints on production within agriculture, at a three dimensional spatial scale, has been limited (Bennett et al., 2015a; Bishop et al., 2015; Gray et al., 2015). This is largely due to the lack of spatial data collection pertaining to soil structure, which is known to be highly influential on the nutrient and water dynamics (Arthur et al., 2013; Bennett et al., 2015a; Quirk, 2001) that control yield. Therefore, obtaining a relationship between soil structural characteristics and crop yield is desirable to allow the exploration of soil factors contributing to yield variability. Furthermore, this will provide capability to provide insight into action management plans for management of soil constraint variability as feasible. Furthermore, such a relationship allows investigation into the potential yield effects of a soil amelioration exercise by simulating the effects of a chemical or mechanical induced structural change. This would assist better establishment of ameliorative management value propositions for constraint management.

The spatially variable nature of soil structure impedes the ability to formulate a relationship with crop yield. Remote and proximal sensing frameworks provide a means to rapidly capture aspects of this variability (Viscarra Rossel et al., 2017; Viscarra Rossel et al., 2010a) although it is well demonstrated that an in-field approach to proximal soil sensing (PSS) is still required (Lobsey and Viscarra Rossel, 2016; Lobsey et al., 2017; Roudier et al., 2017)These systems are not yet commercially adopted to a level that the service is readily available to agricultural operations. Given the limited availability of technology for accurately capturing soil structural variability, there remains an immediate requirement to obtain direct measurements of soil condition with that will augment remote sensing and proximal sensing systems in the future. However, the perceived costs of this direct measurement, coupled with the current limited ability to leverage the data investment in terms of useful on-farm action (e.g. how to implement change in management, or amelioration, based on this data), causes a reluctance in growers and agricultural advisors to collect soil data (Bennett and Cattle, 2013; Bennett and Cattle, 2014; Lobry de Bruyn, 2019). The expense of laboratory soil analyses is largely linked to the volume of laboratory throughput, which is driven by demand, and becomes stymied due to the lack of perceived on-farm data value. To overcome circularity in this argument, a value proposition is required, highlighting the importance and usefulness of site-

specific soil structural datasets in the diagnosis of soil constraints for improved management. This requires soil constraint-yield interactions to be well understood.

Soil-crop interactions are known to be highly complex and non-linear in nature (Dai et al., 2011; Park et al., 2005), meaning they are difficult to empirically model. There exists numerous point-based biophysical and empirical models, such as APSIM (Keating et al., 2003a) and DSSAT (Jones et al., 2003), that attempt to model these interactions. These have traditionally been restricted to soil-nutrient and soil-moisture dynamics, failing to incorporate soil chemical and structural interrelationships (Bennett et al., 2019; Roberton and Bennett, 2017) that are known to be highly influential on the dynamics of soil function and crop performance. Furthermore, such empirical models do not perform well for environments that are highly complex and non-linear in nature (Dai et al., 2011). Resources to manage soil-plant relationships within highly complex and non-linear environments are required. I contend that understanding the soil structural properties defining the dynamics of soil water and nutrients will facilitate the evolution of biophysical/ empirical point based approaches to spatially comprehensive ones.

Non-linear machine learning approaches (referred to here as NLML) offer improved suitability for modelling soil-crop interactions over traditional linear or mechanistic approaches, due to their ability to identify and model complex patterns in data (White, 1989). This stems from the fact that NLML methods do not require underlying assumptions pertaining to data structure (e.g. linearity), which is required at low training data volumes and may not accurately represent the system being modelled. Instead, NLML approaches aim to find the structure during model training. However, without the assumption of an underlying structure having been provided, NLML models create greater bias towards the training examples with an increased sensitivity to noise in the data. This presents an overfitting problem, particularly at low training data volumes where noise is more influential. The ability to identify non-linear relationships within a system thus comes at a cost of increased data requirement, the size of which is largely dependent on the system being investigated, and its inherent variability. Therefore, the superiority of linear or non-linear approaches is dependent on the training data availability of a specific modelling problem. In terms of the diagnosis of soil constraints using soil structural data, this presents a requirement for both approaches to be compared for the given modelling problem.

Machine learning (ML) development in the context of spatial crop prediction has previously been focused on using plant, climate, remotely sensed and proximally sensed data as predictor variables (González Sánchez et al., 2014; Khazaei et al., 2008; Kitchen et al., 2003; Nari and Yang-Won, 2016; Panda et al., 2010; Robinson et al., 2009; Ruß, 2009), as they are comparatively cheap to collect at large spatial and temporal scales. These approaches offer an increased training data volume and therefore present a suitable modelling problem for NLML methods. Whilst these approaches may identify spatial trends in yield performance, they cannot accurately diagnose the cause of this variation, as soil data rarely exists as the same spatial resolution. This presents a significant hurdle in the ability to rapidly explore soil-crop interactions for the purpose of constraint diagnosis using NLML.

Previous attempts towards yield prediction based on directly measured soil attributes have focused on the utilisation of comparatively small training datasets, with little attempt to investigate the risk of overfitting and appropriateness of a NLML approach for the given problem (Dai et al., 2011; Drummond et al., 1998; Irmak et al., 2006; Liu et al., 2001; Niedbała, 2019; Pantazi et al., 2016; Park et al., 2005). Whilst Dai et al. (2011) achieved better model performance using an ANN approach over a linear approach using a training dataset size of 108 observations with 10 variables, Park et al. (2005) achieved improved performance using a linear approach with a training dataset of 720 observations and 22 variables. Therefore, the data volume required for NLML approaches to prevail over linear approaches is problem-specific, and any attempt to investigate the merit of NLML approaches for a given problem requires direct comparison against linear methods.

Additionally, the high-dimensional nature of soil data (i.e. large number of variables) further impedes the success of NLML approaches in data limiting environments, as the resulting datasets often occupy low training observations with a large number of variables (Dai et al., 2011; Irmak et al., 2006; Liu et al., 2001; Niedbała, 2019; Park et al., 2005). Whilst merit exists in applying digital soil mapping (DSM) approaches to augment the training dataset (Pantazi et al., 2016), data dimensionality reduction techniques, such as principal component analysis (PCA), improve model convergence by creating a less convoluted feature space. PCA improves both the speed of model convergence and likelihood of convergence to a global optimal solution. Whilst PCA is widely accepted to improve NLML convergence, its adoption in the literature pertaining to soil-based yield prediction has been limited, and therefore the merit of its use in this context should be investigated.

Differences between training and validation R^2 values have been widely used in the literature as the metric to assess model fit for NLML approaches. Whilst appropriate for linear approaches where overfitting can be directly managed and tested, this metric only provides some insight into the quality of fit, and cannot indicate whether the true soil-yield relationships have been identified, of even if the model is overfitted to the given dataset. In an attempt to optimise yield, Liu et al. (2001) interrogated their model beyond the R^2 metric by manually adjusting input variables and comparing the yield response against known trends. This approach provides a pragmatic estimate of model robustness, providing the adjusted variables remain within the range of that presented in the training observations. Hence, this approach of artificially presenting new information to the model may provide further insight into model performance, but is not usually provided for NLML models relating soil variability to yield (Drummond et al., 1998; Niedbała, 2019; Park et al., 2005).

The applied yield prediction models in the literature are largely limited by the failure to capture subsurface soil-yield interactions. Subsurface conditions are known to be highly influential on crop performance (Price, 2010), and can vary independent of surface conditions. Failure to capture these subsurface relationships provides limited ability to identify yield limiting constraints, which may only be present below the surface layers. Therefore, to use NLML methods to determine the underlying structure, the models must be presented with the data that defines the global structure adequately.

The appropriateness of NLML techniques for yield prediction requires more judicious assessment, as model overfitting has not normally been addressed, nor have linear approaches been used as a baseline for comparison. The effects of training on model convergence has not been thoroughly investigated using high dimensional data sets with few observations for the purpose of yield prediction. Furthermore, there has been little attempt to develop prediction models outside of a single cropping season to achieve a temporally stable locally calibrated model to observe crop responses under various climatic conditions. Therefore, this chapter aims to:

1) Develop season-specific spatial yield prediction models using linear and non-linear approaches
2) Investigate the ability to develop a temporally stable locally calibrated yield prediction model

With the objectives of:

3) Identify the effect of dataset size on model convergence during training
4) Investigate the effect of employing PCA as a data pre-processing technique to reduce data dimensionality
5) Investigate model quality and overfitting beyond the R^2 metric

5.2. Materials and methods

5.2.1. Directly measured soil dataset

The dataset used for yield prediction was collected from a 108 ha dryland cropping paddock in the Warren district of the Macquarie Valley, central NSW. Samples were collected on a 60 m grid at depth increments of 0–10 cm, 10–20 cm, 20–40 cm and 50–60 cm, providing a total of 1200 samples across 300 grid locations. Samples were analysed for soil properties, as presented in Table 5.1. Wheat yield data was obtained for the 2013, 2015 and 2016 winter cropping seasons, the values of which were co-located to the sampling locations using ordinary kriging (OK) within the *automap* package in the R programming environment (Hiemstra and Hiemstra, 2013). For each observation, a total of 9 soil structural features at 4 depths (total of 36 features) were used to describe yield.

Table 5.1. Statistics of measured soil properties

Depth	***Statistic***	***pH***	***EC (ds/m))***	***Clay (%)***	***Silt (%)***	***Sand (%)***	***BD (m/m³***	***K** [cmol(+)/kg]*	***CEC** [cmol(+)/kg]*	***ESP***
0–10 cm	Min	5.27	0.04	3.75	15	10	1.18	0.43	5.78	0.03
	Max	9.15	0.29	71.3	52.5	65	1.83	2.66	38.28	20.86
	Average	6.58	0.10	39.5	28.93	31.56	1.47	1.42	16.55	4.01
	SD	0.64	0.04	10.3	6.52	8.12	0.11	0.39	6.39	3.17
10–20 cm	Min	5.98	0.03	8.75	3.75	13.75	1.37	0.21	7.64	0.13
	Max	9.23	0.37	72.5	45	63.75	1.84	3.51	39.21	26.21
	Average	7.52	0.09	47.8	25.55	26.61	1.61	0.9	23.31	5.32
	SD	0.68	0.04	9.24	5.85	7.64	0.08	0.4	6.16	3.75
20–40 cm	Min	6.55	0.04	20.0	6.25	8.75	1.01	0.2	10.07	0.05
	Max	9.45	0.49	73.8	42.5	55	1.85	2.98	66.5	30.33
	Average	8.23	0.15	50.4	25.62	23.99	1.64	0.73	28.24	7.36
	SD	0.57	0.07	6.98	5.85	6.16	0.08	0.36	5.46	4.69
40–60 cm	Min	5.98	0.06	20.0	3.75	11.25	1.1	0.2	11.05	0.14
	Max	9.65	1.65	67.5	47.5	51.25	1.91	2.39	41.98	34
	Average	8.72	0.24	50.1	25.84	24.07	1.68	0.59	29.57	10.48
	SD	0.56	0.15	6.92	6.28	6.18	0.07	0.31	4.53	5.9

5.2.2. Prediction methods

A brief description of the prediction methods used is given blow

5.2.2.A. Multiple linear regression

Multiple linear regression (MLR) is one of the simplest ML models available and assumes linear dependency between predictor variables and outputs. MLR is formulated as:

$$h_\theta(x_i) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots \theta_m x_m \qquad \text{Equation 5.1.}$$

where θ_m is the coefficient for the predictor variable x_m.

Gradient descent was employed as the training algorithm to iteratively tune the predictor coefficients such that the cost function (Equation 5.2) was minimised. Whilst gradient descent provides a versatile learning method (Qian, 1999), it is prone to finding locally optimal solutions, which are dependent on the location within the feature space from where the model is initialised. To assess the stability of model convergence to an optimal solution, all models were run a total of 50 times to achieve a mean model performance and standard deviation of model predictions.

$$\underset{\theta_1 \ldots \theta_m}{minimise} \quad \frac{1}{2m}\sum_{i=1}^{n}(h_\theta(x_i) - y_i)^2 \qquad \text{Equation 5.2.}$$

5.2.2.B. Regression Trees

Regression trees are a rule-based NLML approach that combines a decision tree model with MLR for the purpose of modelling continuous data. Data is first partitioned using a decision tree model, after which a MLR is fitted at each of the terminal nodes. Cubist (Kuhn et al., 2012) was adopted as the regression tree approach in this chapter. Cubist is based on the widely adopted Quilan M5 model tree (Quinlan, 1992), and has proven ability for modelling in non-linear systems (Malone et al., 2014).

5.2.2.C. Artificial Neural Networks

Artificial Neural Networks (ANNs) are a widely adopted NLML technique used for classification and regression based problems. ANNs consist of layers of parallel processing elements, referred to as neurons, that are connected by sets of weights and biases that represent the level of dependency between neurons. ANNs learn patterns in data by adjusting the sets of weights and biases via forward propagating input variables and back propagating errors. This is achieved by employing the gradient descent algorithm which iteratively tunes network weights and biases with the aim to minimise the cost of the model (i.e. difference between predicted and observed values). This describes the feed-forward backpropagation (FFBP) structure which is employed in this chapter, due to its proven ability for regression in non-linear environments (Haykin and Network, 2004). A three layer network was adopted consisting of a single input layer, hidden layer and output layer.

5.2.2.D. Support Vector Machines

Support Vector Machines (SVMs) differ from other NLML approaches as they employ a Structural Risk Minimisation (SRM) technique that not only ensures a global optimum in the solution (Cristianini and Shawe-Taylor, 2000), but also reduces the likelihood of overfitting, due to the focus of minimising a bound on a risk function, as opposed to minimising training error (Karimi et al., 2008). Whilst SVMs are also commonly referred to as support vector regression (SVR) for the cases of regression-based problems, the notation of SVM will be adopted here. SVMs aim to locate the linear boundaries of the data such that predicted values of the model *f(x)* deviate no greater than distance ε from the observation. This is achieved for non-linear data by application of a kernel that transforms the data into a higher dimensional

feature space, such that it becomes linearly bounded. The radial base kernel function (RBF) was adopted here, due to there being fewer numerical difficulties in comparison with other kernels (Hsu et al., 2003), as only a single hyperparameter (gamma) is required to be specified (Li et al., 2008). A gamma value of 1 was adopted, as suggested by (Üstün et al., 2005).

5.2.3. Principal Component Analysis

NLML models are often susceptible to convergence at a local optimum solution or increased computational times when learning from a dataset occupying a large number of variables and a limited number of training observations. Principal component analysis (PCA) is a pre-processing method employed to reduce such data dimensionality (also called feature reduction) and improve modelling performance. PCA reduces dimensionality by transforming possibly correlated variables into a subset of uncorrelated variables, or principal components (PCs). PCs are found by computing the set of eigenvectors of the covariance matrix (Equation 5.3) for the dataset.

$$\Sigma = \frac{1}{n}\sum_{i=1}^{m}(x^i)(x^i)^T \qquad \text{Equation 5.3.}$$

where n = number of observations, m = number of predictor variables. PCA is investigated in this chapter to determine if dimensionality reduction is beneficial for the applied dataset.

5.2.4. Feature Scaling

Feature scaling is a pertinent data pre-processing method employed to standardise the range of predictor variables such that they are presented to the network at a comparable variance. Feature scaling improves model performance and reduces the likelihood of convergence to a local optima. Mean normalisation is used in this chapter as the feature scaling method for all datasets and is presented as follows:

$$z = \frac{x_i^j - \mu^j}{\sigma^j} \qquad \text{Equation 5.4.}$$

where z is the new scaled value for x, x is the ith observation of the jth feature, μ is the mean value of the jth feature, and σ is the range of the jth feature.

5.2.5. *Model Assessment*

5.2.5.A. *Effect of training data size on model convergence*

This was assessed by applying datasets of size 300 and 29,978 respectively to predict yield variability for the 2013, 2015 and 2016 cropping seasons individually. The original directly measured dataset of 300 observations was used to spatially interpolate a dataset of 29,978 observations at a 6 m pixel resolution, *automap* package in the R programming environment (Hiemstra and Hiemstra, 2013). A multiple linear regression, regression tree, ANN and SVM model were fitted to the data to provide a yield function:

$$Yield = f(depth_1, depth_2, depth_3, depth_4) \quad \text{Equation 5.5.}$$

where

$$Depth_i = f(soil\ pH_i, soil\ EC_i, clay\%_i, silt\%_i, sand\%_i, BD_i, CEC_i, ESP_i, K_i) \quad \text{Equation 5.6.}$$

A total of 36 features were used to predict yield as the dependent variable for each observation. The training data was spatially partitioned to assign 80% of the data for training and 20% for validation, to undertake internal validation (i.e. validation within the same field and cropping year). Each model was trained and validated a total of 50 times to observe sensitivity of model convergence to random initialisation of model parameters. All features were scaled using mean normalisation prior to training.

5.2.5.B. *Assessing and reducing model overfitting*

Model overfitting was explored beyond the metric of R^2 difference between training and validation as this does not accurately diagnose the presence of overfitting, especially when the validation data is similar to that of training. Overfitting was assessed for two sets of models, namely i) Models trained using the original 36 features; and ii) Models trained using PCs of the dataset. The latter set of models were trained using the first 8 PCs as features which explained 99% of the variability within the dataset. Overfitting was assessed by identifying the ability of the trained models to identify known trends in crop response to changes in soil condition (i.e. soils with higher ESP generally have lower soil structural stability and therefore lower crop potential). This was achieved by manually adjusting field conditions within the

bounds of the training data and subsequently observing the models' ability to inform practically sensible crop response against established and, hence, expected yield response trends to the manipulated feature. In this case ESP was used to inform a chemical or structural change at a paddock total level, whereby an increase in the ESP was expected to result in a decrease in crop yield (Bennett et al., 2016; Rengasamy et al., 1984). The trained model was tested at ESP values 3-9, to observe the effects on predicted crop yield and the model's ability to detect sensible trends, irrespective of R^2. Whilst it is not possible to assess the predictive accuracy of this new simulated data, the ability of the models to identify known trends can be identified, therefore providing insight into the quality of model fit.

5.2.5.C. Temporally stable locally calibrated yield prediction model

The 3 independent wheat cropping years were used to investigate the potential of developing a temporally stable yield prediction model, where yield predictions could be estimated under different weather scenarios. This would allow for interactions between weather, soil structure and crop response to be investigated in the temporal domain. This was investigated using a leave-one-out approach, whereby two cropping years were combined for training and the model validated on the third independent year as per Table 5.2.. Input features used in model development were reduced from 36 soil variables to 8 PCs. The simulations were as per.

Table 5.2. Simulations to investigate the development of a locally calibrated yield prediction model

Simulation	*Prediction year*	*Training year*
1	2013	2015 & 2016
2	2015	2013 & 2016
3	2016	2013 & 2015

Seasonal rainfall for each season were used as features in conjunction with the available soil structural information. Seasonal rainfall was summed from October of the previous year to September of the cropping year to account for filling of the profile during the summer fallow. This was obtained using data obtained from a farm-based digital weather station. Rainfall for the 2013, 2015 and 2016 cropping season was 380 mm, 500 mm and 691 mm respectively.

5.3. Results

In general all models provided reasonable prediction of single year yield variability, but performed poorly in the creation of a temporally stable, localised model for multiple years of

yield. Further interrogation of the single year yield prediction models also identified that model behavior varied between the models. This indicates a requirement for soil science domain knowledge in interrogating model performance, especially when seeking to inform the specific effect of model components on yield outcome.

5.3.1. Season-specific spatial yield prediction models

5.3.1.A. The effect of PCA

The MLR, Cubist and ANN models achieved the greatest R^2 for both training and validation using the original 36 feature dataset (Figure 5.1) (i.e. no data dimensionality reduction completed). The difference between R^2 of training and validation was also minimal. This would suggest that the models are well fitted. However, when performance of these models was tested beyond this metric, by exploring known soil-crop relationships, it was found that they were not well generalized, and did not detect known trends between ESP and yield (Figure 5.2). Predicted soil-crop interactions are only shown here for the 2013 season to reduce repetition within the results section. Total site yield should decrease with increases in ESP (Bennett et al., 2016; Rengasamy et al., 1984), however, the predicted trends of the developed models were often nonsensical, with predictions frequently indicating a yield increase with increases in ESP (e.g. ANN yield predictions based on ESP changes in the 20–40 cm and 40–60 cm depth layers). Furthermore, the identified trends were highly inconsistent between models and depths, with the models identifying a yield increase due to an increased ESP in one depth layer, and subsequently identifying a yield decrease due to an increased ESP change in a second depth layer (e.g. ANN in the 10–20 cm and 20–40 cm depth layers). This suggests that whilst a superior R^2 was obtained for these models, they were severely overfitted to the data and were not capable of detecting the generalised trends.

Whilst employing PCA as a data dimensionality technique to decrease the dataset to 8 features consequently reduced the R^2 of training and validation for the MLR, Cubist and ANN models (Figure 5.1), the developed models were better generalised when assessing ESP and yield relationships (Figure 5.2). In general, all models were able to detect a decrease in crop yield with increasing ESP, however the rate at which yield decreased was inconsistent between models. This suggests that the PCA-based models are better fitted to the data, despite a reduced R^2 of prediction. Interesting, the Cubist model detected a yield increase at an ESP of 5 for all depth layers, whereby yield decreased thereafter, suggesting that ESP may not be the driver of yield at low values.

When considering the R^2 metric, the SVM model achieved substantially better predictive performance after applying PCA, with values approximating 0.99 for training and validation for all seasons. R^2 values of this magnitude present a large concern for model overfitting and are unrealistic in environmental modelling, where the natural variation and uncertainty is high. Whilst acceptable ESP-yield trends were identified for the SVM in 0–10 cm and 10–20 cm surface layers, yield increases were predicted at ESP>6 in the 20–40 and 40–60 cm subsurface layers. This suggests that even with PCA, the SVM was not well generalised, particularly for the sub-surface layers.

Although testing the models against known soil-yield trends (i.e. adjusting ESP) provided insight toward the sensitivity of the methods to overfitting, it is worthwhile noting this approach remains limited. Whilst the bounds of ESP values were kept consistent with that observed within the dataset, there is no guarantee that these adjustments remained within the bounds of the multivariate space, therefore attempts to predict outside of the multivariate training data space may have occurred. This again highlights the difficulty in interpreting ML models to predict complex soil-crop interactions and reinforces that better interpretability metrics are required.

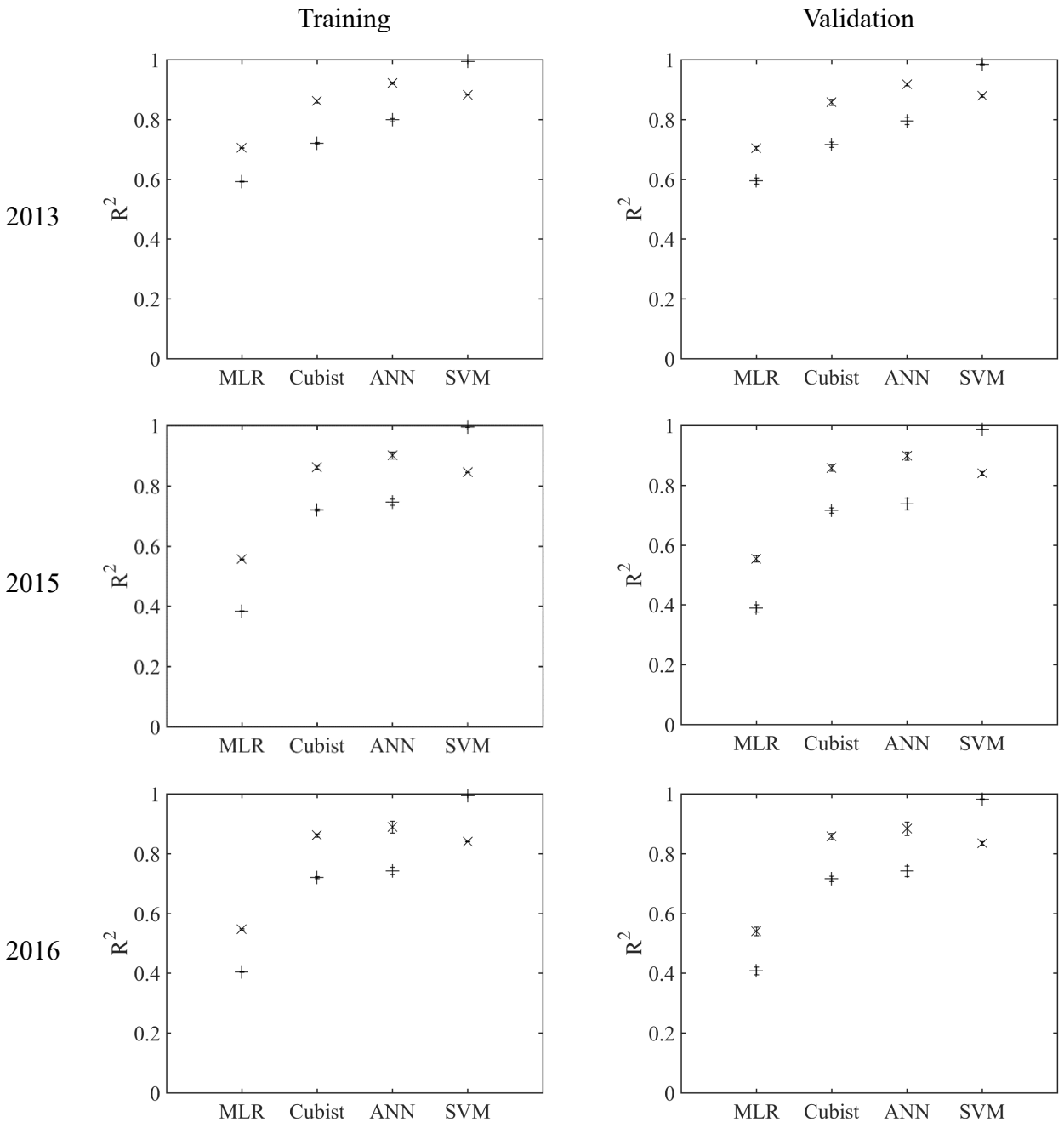


Figure 5.1. Training and validation results for model development using the original normalised dataset with 36 features (left) and using the dataset reduced to 8 PCs (right) for 3 wheat cropping years. Validation represents internal validation, where each model is validated using data within the same field and cropping year.

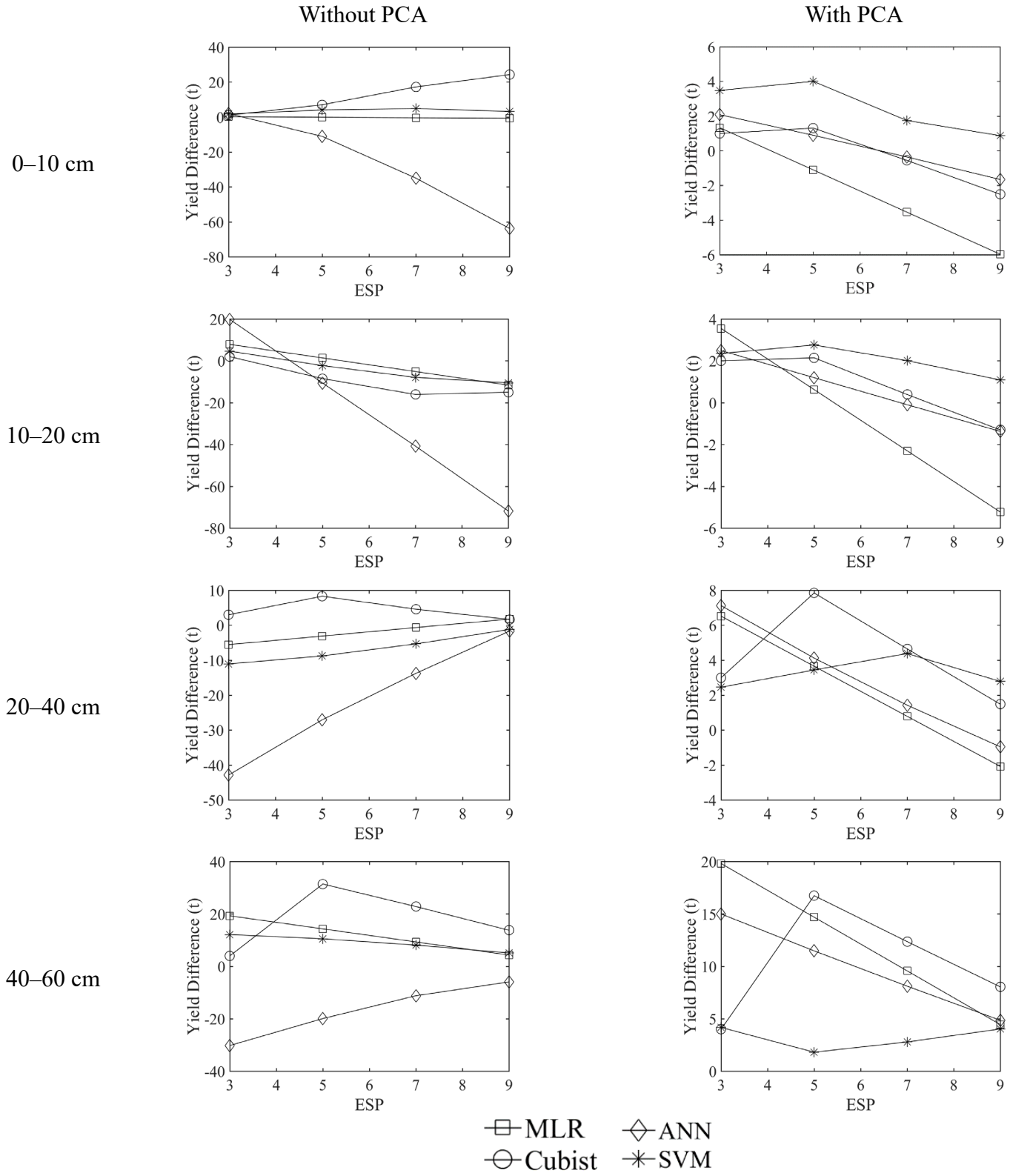


Figure 5.2. Paddock total yield predictions using the MLR, Cubist, ANN and SVM models developed using the original 36 feature dataset (left) and the dataset reduced to 8 PCs (right) for the 2013 cropping season

5.3.1.B. The effect of training size on model convergence

The size of the training dataset did not greatly affect the reliability of model convergence for the MLR, Cubist and SVM models, as the difference in training R^2 values between iterations of the model was minimal (Figure 5.3). This suggests that model convergence was not sensitive to the random initialisation of search parameters in finding a global optimal solution in the training data. However, highly unstable predictions were observed for the reduced dataset size of 300, as training and validation R^2 varied greatly between model iterations. This suggests that whilst a global optimal solution was achieved in each training iteration, the models were either overfitted to the data presented for training, or the inherent variability in the dataset was of a magnitude which resulted in vast differences between the partitioned training and validation datasets.

At low data densities, the ANN was highly unstable between model iterations, for both training and validation, suggesting that the model often converged to local optimal solutions and was sensitive to the random initialisation of search parameters, despite PCA being employed. The ANN was the most unstable of all the methods and displayed the lowest predictive performance during validation for a dataset size of 300. This however was overcome at an increased dataset size of 29,978 observations, which likely allowed for reliable convergence to global optimal solutions between iterations.

An increased dataset size of 29,978 observations resulted in better model performance for all methods, when considering the R^2 metric. The MLR consistently performed poorly, with greater predictive accuracies being obtained by the Cubist, ANN and SVM models, respectively. In general, the ANN produced only slightly better R^2 results for training and validation, as compared to the Cubist model. R^2 for the SVM remained high (>0.99) for both training and validation, suggesting overfitting was occurring.

5.3.2. Temporally stable localized calibration

Of the 4 models applied, none were capable of predicting yield variability within a single independent year, using 2 years of soil, yield and weather data for training. This can be observed in Figure 5.4 to Figure 5.6 which display the training and testing performance of the models for the 2013, 2015 and 2016 wheat cropping test years. Whilst all models were able to achieve acceptable to high R^2 values during training, testing performance of the models in independent years decreased greatly and displayed little to no correlation with the observed testing data. For the 2015 and 2016 prediction years, the MLR, Cubist and ANN models did

not accurately represent the range of yield variability observed at the site, with low yields being over predicted, and high yields being under predicted. This is seen by the horizontal linear distribution within the testing data for these years (Figure 5.5 and Figure 5.6).

Interestingly, the testing results obtained for the SVM model across all years displayed a distinct horizontal linear pattern in the predictions, suggesting a fundamental issue with SVM predictions when predicting outside of the bounds of training data. The SVM R^2 values for training were high (0.99) for all years, signifying the model over fitted to the training observations in all instances. The SVM model was not able to predict variability within the testing data, and instead, provided a default value for each observation which was equal to the mean of the training data set. Generally, the NLML methods struggled to a higher degree of sensitivity when attempting to extrapolate from the limited training dataset.

Table 5.3. Statistics of training and testing datasets for temporally-stable yield model development

	2013		*2015*		*2016*	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Train observations	3.34	1.44	2.66	1.69	2.05	1.051
Testing observations	1.37	0.339	2.74	1.07	3.95	1.5
MLR test	1.97	0.73	2.37	0.55	4.91	0.434
Cubist test	2.12	0.71	1.35	0.261	4.91	1.13
ANN test	3.61	1.05	2.06	1.326	0.51	0.98
SVM test	3.35	0	2.775	0	2.06	0

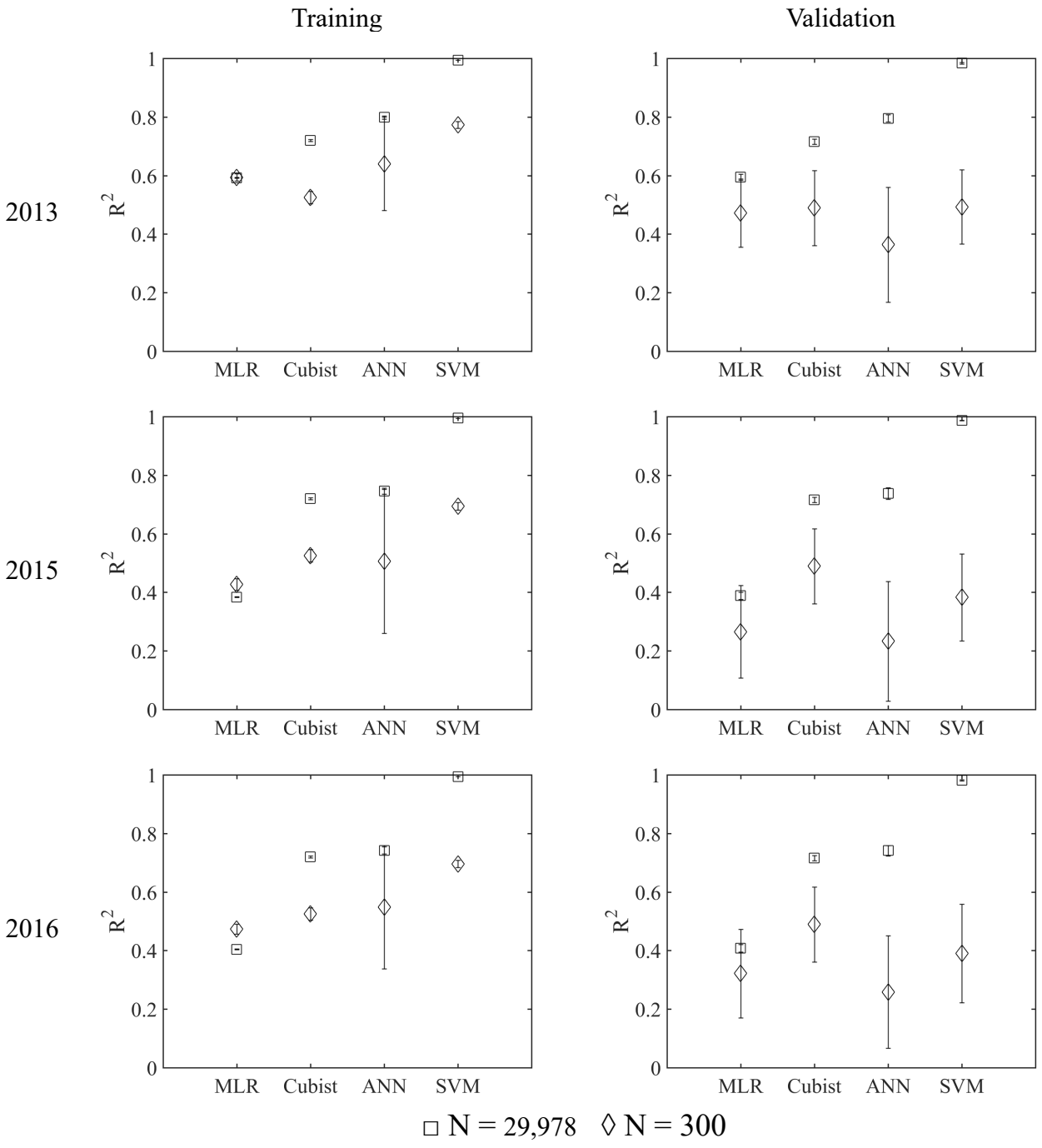


Figure 5.3. Mean training and validation results for MLR, cubist, ANN and SVM yield prediction models for the 2013, 2015 and 2016 wheat cropping seasons using dataset densities of 300 and 29,978. Error bars represent 1 standard deviation of the model results of 50 iterations

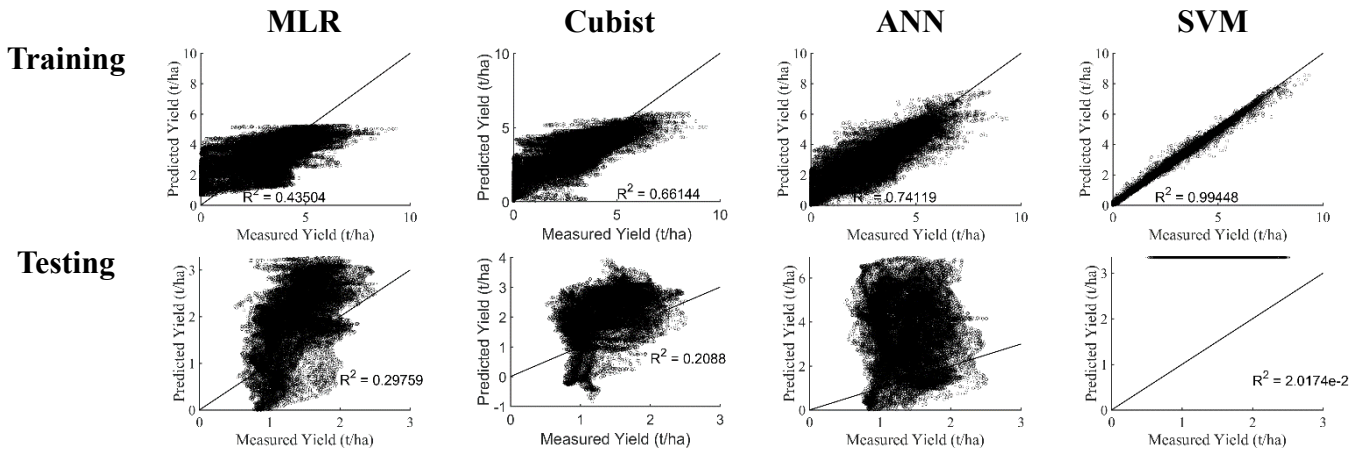


Figure 5.4. Training and validation results for simulation 1 – predicting yield variability in the 2013 cropping season using 2015 and 2016 and training years

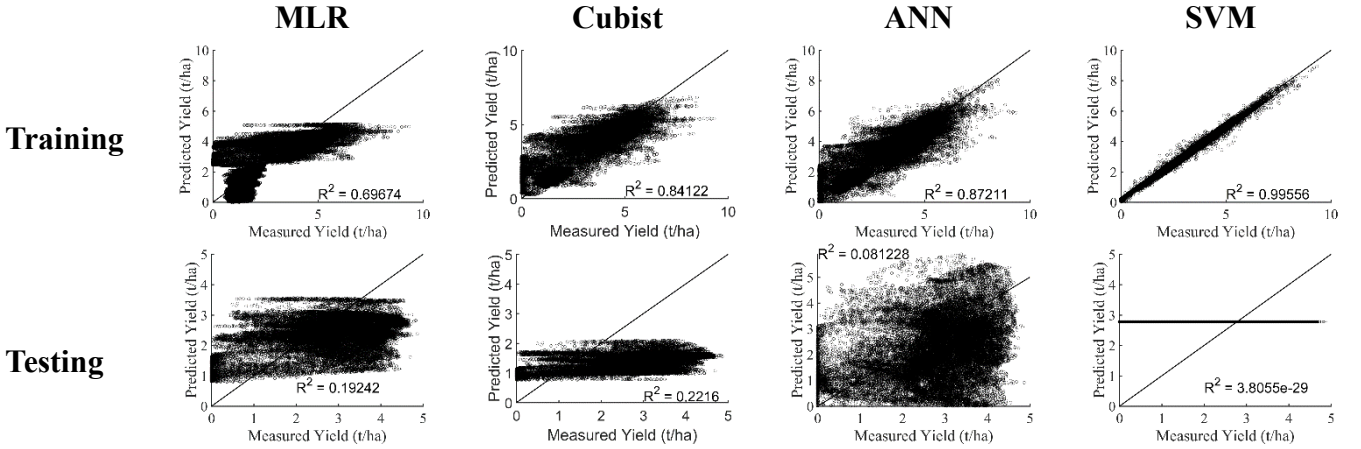


Figure 5.5. Training and validation results for simulation 2 – predicting yield variability in the 2015 cropping season using 2013 and 2016 and training years

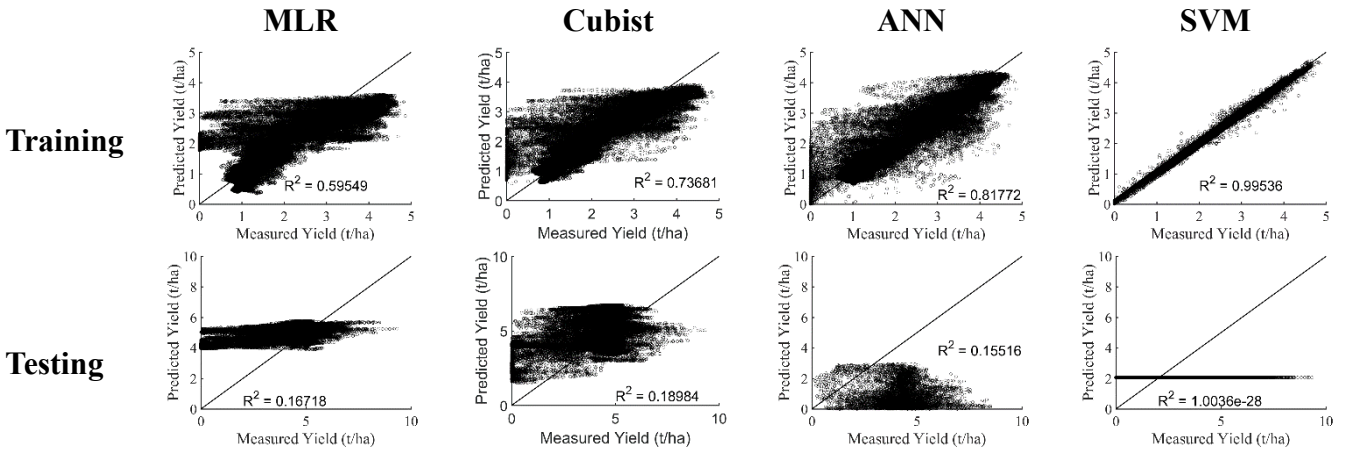


Figure 5.6. Training and validation results for simulation 2 – predicting yield variability in the 2016 cropping season using 2013 and 2015 and training years

5.4. Discussion

5.4.1. From single season to temporally stable predictive models

Whilst 300 directly measured soil cores were sufficient to train a yield prediction model to explain spatial yield variability within single seasons, inference could not be extended to independent years, meaning a generalised temporally stable yield prediction model could not be developed. The ability to achieve this is inhibited by the lack of available temporal information, which is reduced to only two years of training cases in order to test within a third, independent year. Whilst this contrasts the findings Irmak et al. (2006) of who achieved R^2=0.57 within a third, independent year, they used a site of 20 ha and ~0.26 samples/ha. We contend that the 108 ha commercially operated site investigated in this study was represented by a greater degree of inherent variability to data density ratio. For single year predictions, validation performance was similar to that found by Dai et al. (2011), Drummond et al. (1998), Irmak et al. (2006), Liu et al. (2001) and Pantazi et al. (2016) when only considering R^2 as a measure of fit.

The lack of temporal training observations inhibits model generalisation as the influence of noise attributed to variables unrelated to soil structure or weather (e.g. pest, disease and sub-field weather variation) is increased. This can be overcome by capturing these factors and incorporating them as dependent variables within the model, or by increasing the volume of temporal observations. Whilst attempts have been made to spatially detect the influence of pest and disease on yield at the sub-field scale (Faithpraise et al., 2013; Sankaran et al., 2010), this technology is not currently mature enough to accurately account for their affects, however, this should remain a future focus. Accounting for the noise produced by unexplained variables therefore requires increased temporal observations in the form of seasonal yield and weather data. However, future work is required to investigate if this approach is achievable within a reasonable time period. Furthermore, the presence of only 2 observations in the weather-yield feature plane does not allow nonlinear trends to be identified.

Without a temporally stable generalised yield prediction model, future yield responses due to a system change cannot be accurately assessed. Development of such a model would be advantageous to aid in the identification of yield limiting soil constraints and to subsequently estimate the yield effects of a chemically, or structurally, induced change via application of an ameliorant. This would allow for economical assessment to optimise on-farm soil investments. It remains possible to achieve such assessments using single-season models, although the

outcome is biased to the training year/s, which may not necessarily be a representative season. Furthermore, single season models do not allow for the effects of climate uncertainty to be assessed against the yield response due to an induced change. Until sufficient data is collected to improve generalisation of temporal yield models, single-season models may be applied to estimate yield responses, due to a system change. A level of caution must be applied when using these models as there is no guarantee the results obtained are temporally representative at the site.

Interestingly, the SVM model exhibited unique behavior when predicting in the independent years, as the value of all observations was defaulted to the mean of that observed in training. It is hypothesized that this is caused by the testing data occupying values beyond the bounds of the slack variables which represent the outer limits of model errors (Vapnik, 1995). This may occur when the SVM is overfitted to a dataset and new data is provided that is outside of the bounds of training; i.e. the new rainfall feature occupies a value that is dissimilar to that observed for the feature during training. Further investigation is required to confirm this hypothesis and determine whether it is a failing of the underlying theory, or the specific modelling toolbox employed in the Matlab® programming environment.

5.4.2. Improving generalisation of yield prediction models

Achieving satisfactory generalisation for ML yield prediction models is pertinent to ensure model predictions are representative of the soil-yield relationships. This requires consideration towards avoiding both overfitting and convergence to local optimal solutions during training. Overfitting and local convergence were both observed in this work, which reduced the generalisation of the development models. This was found to be influenced by the number of observations in the training dataset and the dimensionality of this data. These two phenomena are discussed in the following sections.

5.4.2.A. The effect of data size on generalisation

Models trained using the original 300 data observations produced poor generalisation results. This was observed by a large reduction in R^2 values between training and validation simulations, indicating that overfitting was present during training. This effect was greatest for the nonlinear ANN and SVM models, with only a small reduction in R^2 being observed for the MLR and Cubist models for all prediction years. This signifies that for the given modelling problem, 300 observations was not sufficient for pure nonlinear methods to find structure within the data and prevail over linear approaches. Whilst Cubist is considered a nonlinear

regression model (Malone et al., 2014), its hybrid approach allows for the key benefits of linear methods in data limiting environments to be utilised. This is achieved by using a non-linear decision tree model to initially partition the data into discrete subsets which MLR models are subsequently applied to, therefore resulting in smaller errors than that of a single regression applied to the entire dataset (Quinlan, 1992). At a sample size of 300 observations, the applied Cubist model achieved the greatest validation R^2 for all modelling years, highlighting its advantages over both standard linear approaches and advanced non-linear approaches in data limiting environments.

At the reduced training data volume of 300, the MLR and Cubist model were able to achieve better generalisation over nonlinear ANN and SVM approaches, due to the linear structural assumptions that they apply. These assumptions are often required in data limiting environments to reduce the influence of training data noise on model convergence. Whilst these assumptions impose limitations when modelling non-linear relationships, a compromise is required that simultaneously assesses the error of over-simplifying the modelling problem versus the error of overfitting. Consider the 2 dimensional problem presented in Figure 5.7 which represents linear and nonlinear model fitting for two data densities. At low training data densities, the nonlinear model over-fits to the data in an attempt to minimise error. Whilst the linear model is not well fitted to the true relationship, its generalisation is better than that of the nonlinear model, and so a comprise is required. Furthermore, the lack of structural assumptions at low densities biases nonlinear methods to the observations provided in training, meaning that new training observations in a second iteration may result in a vastly different model, despite the R^2 values of training being equivalent. As sample density increases however, the influence of noise in the training data is reduced and better generalisation may be found using a nonlinear approach.

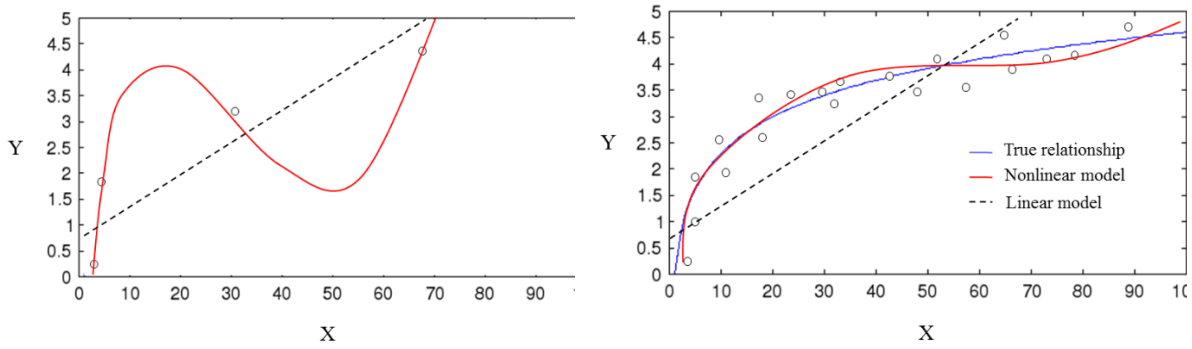


Figure 5.7. Illustration of a 2-dimesnional modelling problem. Linear and nonlinear model fitted at a training density of N = 4 (left) and N = 20 (right). The linear model is more generalised compared with the overfitted nonlinear model at N=4. Nonlinear generalisation however increases as N increases.

Generalisation was improved for all models by using data augmentation to spatially interpolate direct measurements. Interpolation artificially created more observations to improve model convergence. These new observations were generated using geostatistical theory (Burgess and Webster, 1980a; Burgess and Webster, 1980b) which states soil some location *(x,y)* is dependent on the geographic location *(x,y)* and soil at neighboring locations *(x+u, y+v)* (McBratney et al., 2003). OK was used, which has been shown to accurately approximate soil properties at unknown locations by augmenting the directly measured data at known locations (Burgess and Webster, 1980a; Burgess and Webster, 1980b). Data augmentation benefits nonlinear learning in small datasets by artificially increasing the sample site to reduce the risk of overfitting and improve generalisation (Perez and Wang, 2017; Santoro et al., 2016). Increasing data reduces the bias towards the training observations by including more observations with a larger variance. Therefore, for nonlinear modelling problems, it can be advantageous to seek more, lower quality data over fewer, precise observations. Data augmentation is more commonly used for image classification problems (Bargoti and Underwood, 2017; Fawzi et al., 2016; Krizhevsky et al., 2012) by rotating training images, however can be applied to regression problems when some prior knowledge regarding the distribution of features exists. Geostatistical theory represents this prior knowledge for spatial applications by allowing for robust generation of new observations that are representative of the original data. Spatial interpolation methods such as OK are therefore useful to provide data augmentation to improve generalisation of NLML methods in spatial prediction applications.

5.4.2.B. The effect of data dimension on generalisation

Generalisation of all models was significantly improved by employing PCA as a data dimensionality reduction technique, which removed features that were not well correlated to yield an instead presented noise within the data (Demšar et al., 2013). Importantly, this could only be assessed by manually adjusting model variables and observing the yield response against known soil-crop relationships. The difference in R^2 between training and validation provided misleading confidence toward generalisation, as without PCA, all models achieved high R^2 during training and validation, however, when presented new information, these models did not identify sensible soil-yield trends. On the contrary, models developed from the reduced feature dataset identified these trends, despite lower quality R^2 being achieved. This demonstrates the risk of over-relying on the R^2 metric as an indicator of generalisation, as it only indicates how well the model has fitted to the individual observations, not the overriding general trend (Alexander et al., 2015). Therefore, using only this metric can provide an illusion of the confidence of the model to produce accurate and reliable results. This is a key shortfall of model development in the literature (Drummond et al., 1998; Niedbała, 2019; Park et al., 2005) and requires future attention to better assess over fitting and generalisation beyond R^2. This requires better interpretation techniques to be applied.

The effects of poor generalisation within the context of agricultural management can be highly detrimental if the developed yield model is used to identify yield-limiting constraints and subsequently provide amelioration advice. An overfitted model incapable of accurately describing the soil-yield relationships will provide incorrect advice for resource application (e.g. fertilizer, gypsum etc.), which can be economically and agronomically detrimental. Therefore, appropriate feature selection is pertinent in model establishment to reduce the risk of overfitting and negative effects of false model predictions.

5.5. Conclusion

ML approaches were found to accurately describe yield variability within independent seasons using a directly measured soil dataset. Of the methods investigated, the nonlinear Cubist and ANN models prevailed over the MLR and SVM models, with better generalisation being observed. Prediction uncertainty of the ANN was significantly greater than that of the other models at low training densities (i.e. 300), suggesting the model was susceptible to local optima solutions and overfitting. Whilst this uncertainty was reduced at increased training data, and sensible trends were observed when testing yield response against ESP, the black-box

nature of ANNs inhibits model interpretability. It is therefore difficult to confidently conclude generalisation was achieved. Future work should focus on investigating better interpretability methods for ANN assessment. We therefore see Cubist as the superior nonlinear approach for modelling site-specific yield variability based on soil properties.

Whilst season-specific yield models produced accurate predictions, a temporally stable model could not produce reliable results for independent seasons, due to the reduced temporal resolution within the training data. This however can be improved by incorporating additional years of yield data, if available, to account for the effects of temporally variability. This should remain a focus of future work.

Applying OK as a data augmentation technique to artificially increase training data volume greatly improved model convergence and fit, when considering the R^2 metric. Data augmentation can improve generalisation of ML models by reducing the influence of noise within small training sets. Appropriate consideration is however required when augmenting data to ensure the new information is representative of the system being modelled. OK provides a useful method for augmentation in spatial predication problems.

Data dimensionality reduction using PCA was found to be pertinent in achieving appropriate generalisation. The use of PCA reduced the number of training features to 8, which represented 99% of the variability within the original 36 features. Although the models developed using the original 36 features achieved greater R^2 for training and validation, they were not able to reliability identify known soil-yield trends. This highlights the requirement for model interpretability when assessing generalisation, as comparing the R^2 between training and validation can give misleading confidence on the quality of fit. Whilst known soil-yield trends were used here to further assess model performance beyond the R^2 metric, this is often not achievable as no knowledge of system dynamics is known prior. Therefore, future work must consider better interpretation techniques that can provide insight into generalisation for linear and nonlinear crop yield models.

5.6 References

Alexander, D., Tropsha, A., Winkler, D.A., 2015. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. Journal of chemical information and modeling 55(7), 1316-1322.

Arthur, E., Moldrup, P., Schjønning, P., de Jonge, L.W., 2013. Water retention, gas transport, and pore network complexity during short-term regeneration of soil structure. Soil Science Society of America Journal 77(6), 1965-1976.

Bargoti, S., Underwood, J., 2017. Deep fruit detection in orchards, 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 3626-3633.

Bennett, J.M., Cattle, S., 2013. Adoption of soil health improvement strategies by Australian farmers: I. Attitudes, management and extension implications. The Journal of Agricultural Education and Extension 19(4), 407-426.

Bennett, J.M., Cattle, S., 2014. Adoption of soil health improvement strategies by Australian farmers: II. Impediments and incentives. The Journal of Agricultural Education and Extension 20(1), 107-131.

Bennett, J.M., Cattle, S., Singh, B., 2015. The efficacy of lime, gypsum and their combination to ameliorate sodicity in irrigated cropping soils in the Lachlan Valley of New South Wales. Arid Land Research and Management 29(1), 17-40.

Bennett, J.M., Marchuk, A., Raine, S., Dalzell, S., Macfarlane, D., 2016. Managing land application of coal seam water: A field study of land amendment irrigation using saline-sodic and alkaline water on a Red Vertisol. Journal of environmental management 184, 178-185.

Bennett, J.M., Roberton, S., Marchuk, S., Woodhouse, N., Antille, D., Jensen, T., Keller, T., 2019. The soil structural cost of traffic from heavy machinery in Vertisols. Soil and Tillage Research 185, 85-93.

Bishop, T., Horta, A., Karunaratne, S., 2015. Validation of digital soil maps at different spatial supports. Geoderma 241, 238-249.

Burgess, T., Webster, R., 1980a. Optimal interpolation and isarithmic mapping of soil properties: I. The semivariogram and punctual kriging. Journal of soil science 31(2), 315-331.

Burgess, T., Webster, R., 1980b. Optimal interpolation and isarithmic mapping of soil properties: II block kriging. Journal of Soil Science 31(2), 333-341.

Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Dai, X., Huo, Z., Wang, H., 2011. Simulation for response of crop yield to soil moisture and salinity with artificial neural network. Field Crops Research 121(3), 441-449.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., McLoone, S., 2013. Principal component analysis on spatial data: an overview. Annals of the Association of American Geographers 103(1), 106-128.

Drummond, S., Joshi, A., Sudduth, K.A., 1998. Application of neural networks: precision farming, Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on. IEEE, pp. 211-215.

Faithpraise, F., Birch, P., Young, R., Obu, J., Faithpraise, B., Chatwin, C., 2013. Automatic plant pest detection and recognition using k-means clustering algorithm and correspondence filters. International Journal of Advanced Biotechnology and Research 4(2), 189-199.

Fawzi, A., Samulowitz, H., Turaga, D., Frossard, P., 2016. Adaptive data augmentation for image classification, 2016 IEEE International Conference on Image Processing (ICIP). Ieee, pp. 3688-3692.

González Sánchez, A., Frausto Solís, J., Ojeda Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction.

Gray, J.M., Bishop, T.F., Yang, X., 2015. Pragmatic models for the prediction and digital mapping of soil properties in eastern Australia. Soil Research 53(1), 24-42.

Haykin, S., Network, N., 2004. A comprehensive foundation. Neural networks 2(2004), 41.

Hiemstra, P., Hiemstra, M.P., 2013. Package 'automap'. compare 105, 10.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A practical guide to support vector classification.

Irmak, A., Jones, J., Batchelor, W., Irmak, S., Boote, K., Paz, J., 2006. Artificial neural network model as a data analysis tool in precision farming. Transactions of the ASABE 49(6), 2027-2037.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. European journal of agronomy 18(3-4), 235-265.

Karimi, Y., Prasher, S., Madani, A., Kim, S., 2008. Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. Canadian Biosystems Engineering 50(7), 13-20.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N., Meinke, H., Hochman, Z., 2003. An overview of APSIM, a model designed for farming systems simulation. European journal of agronomy 18(3-4), 267-288.

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

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.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, pp. 1097-1105.

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

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis & Prevention 40(4), 1611-1618.

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

Lobry de Bruyn, L., 2019. Learning opportunities: Understanding farmers' soil testing practice through workshop activities to improve extension support for soil health management. Soil Use and Management.

Lobsey, C., Viscarra Rossel, R., 2016. Sensing of soil bulk density for more accurate carbon accounting. European Journal of Soil Science 67(4), 504-513.

Lobsey, C., Viscarra Rossel, R., Roudier, P., Hedley, C., 2017. rs‐local data‐mines information from spectral libraries to improve local calibrations. European journal of soil science 68(6), 840-852.

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232, 34-44.

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

Nari, K., Yang-Won, L., 2016. Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State. Journal of the Korean Society of Surveying Geodesy Photogrammetry and Cartography (34(4)), 383-390.

Niedbała, G., 2019. Simple model based on artificial neural network for early prediction and simulation winter rapeseed yield. Journal of integrative agriculture 18(1), 54-61.

Panda, S.S., Ames, D.P., Panigrahi, S., 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sensing 2(3), 673-696.

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

Park, S., Hwang, C., Vlek, P., 2005. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. Agricultural Systems 85(1), 59-81.

Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Price, P., 2010. Preface: Combating Subsoil Constraints: R&D for the Australian grains industry. Soil Research 48(2), i-iii.

Qian, N., 1999. On the momentum term in gradient descent learning algorithms. Neural networks 12(1), 145-151.

Quinlan, J.R., 1992. Learning with continious classes, Australian Join Conference on Artifical Intelligence, Hobart, Vic, pp. 343-348.

Quirk, J., 2001. The significance of the threshold and turbidity concentrations in relation to sodicity and microstructure. Soil Research 39(6), 1185-1217.

Rengasamy, P., Greene, R., Ford, G., Mehanni, A., 1984. Identification of dispersive behaviour and the management of red-brown earths. Soil Research 22(4), 413-431.

Roberton, S.D., Bennett, J.M., 2017. Efficacy of delaying cotton defoliation to mitigate compaction risk at wet harvest. Crop and Pasture Science 68(5), 466-473.

Robinson, N., Rampant, P., Callinan, A., Rab, M., Fisher, P., 2009. Advances in precision agriculture in south-eastern Australia. II. Spatio-temporal prediction of crop yield using terrain derivatives and proximally sensed data. Crop and Pasture Science 60(9), 859-869.

Roudier, P., Hedley, C., Lobsey, C., Rossel, R.V., Leroux, C., 2017. Evaluation of two methods to eliminate the effect of water from soil vis–NIR spectra for predictions of organic carbon. Geoderma 296, 98-107.

Ruß, G., 2009. Data mining of agricultural yield data: A comparison of regression models, Industrial Conference on Data Mining. Springer, pp. 24-37.

Sankaran, S., Mishra, A., Ehsani, R., Davis, C., 2010. A review of advanced techniques for detecting plant diseases. Computers and Electronics in Agriculture 72(1), 1-13.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T., 2016. Meta-learning with memory-augmented neural networks, International conference on machine learning, pp. 1842-1850.

Üstün, B., Melssen, W.J., Oudenhuijzen, M., Buydens, L.M.C., 2005. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. Analytica Chimica Acta 544(1), 292-305.

Vapnik, V., 1995. The nature of statistical learning theory. J. Wiley & Sons, New York.

Viscarra Rossel, R.A., Lobsey, C.R., Sharman, C., Flick, P., McLachlan, G., 2017. Novel Proximal Sensing for Monitoring Soil Organic C Stocks and Condition. Environmental Science & Technology 51(10), 5630-5641.

Viscarra Rossel, R.A., McBratney, A., Minasny, B., 2010. Proximal soil sensing.

White, H., 1989. Learning in artificial neural networks: A statistical perspective. Neural computation 1(4), 425-464.

6. Towards identifying the soil data investment to economically optimise soil ameliorant recommendations as a function of yield

6.1. Introduction

The cost of soil ameliorants to amend interacting soil constraints presents a significant investment for landholders. Gypsum ($CaSO_4.2H_2O$) and lime ($CaCO_3$) have been documented as $110 and $75 per ton (transported and spread; Bennett et al., 2015a), respectively, with this price fluctuating based on distance to market, product purity, exchange efficiency (affected by rainfall/ irrigation inputs), and agricultural inputs, such as irrigation water (Abbott and McKenzie, 1986; Bennett et al., 2016; Greene and Ford, 1985). Considering the extent of dispersive soils within Australia, both spatially and with depth, gypsum application rates for amelioration would regularly exceed 10 t/ha within the top 20 cm of the soil profile (Doyle and Habraken, 1993; Ford et al., 1993; McKenzie et al., 1993; Naidu et al., 1993; Shaw et al., 1994; Tennant et al., 1992), which equates to an amelioration cost often in excess of $1000 /ha. Given the liklihood of multiple constraints interacting within Australian cropping soils (Dang et al., 2006; Dang et al., 2010) the true cost of amelioration would exceed this estimate, and regularly exceed the per hectare price of land, making the value proposition of homogeneous, full-field ameliorant approaches extremely unattractive. A variable rate approach to soil amelioration, based on the sub-paddock scale distribution of constraints, would facilitate a more tenable approach for economic investment.

In Chapter 4, the minimum number of samples required to produce a spatial constraint map with accuracy comparable to that of a 60x60 m sampling grid was considered. This approach provided the ability to diagnose constraints, but does not address the value proposition of investment in amelioration. The number of soil samples required to optimise an agronomic recommendation must be considered in terms of the cost of sampling, the economic quantification of the ameliorant application error, and the potential return on investment as a function of yield. The accuracy of agronomic recommendations are directly dependent on the ability to map and identify soil conditions. Increasing soil sampling requirements improves the accuracy of spatial prediction, thus subsequently reducing error of the recommendation. However, the costly nature of soil sampling constrains the ability to simply collect more data. An economically optimised point therefore exists where sampling density is minimised, whilst maximising the accuracy of agronomic recommendations. This will be referred to as the minimum dataset.

There has been limited attempt to quantify how spatial prediction error of digital soil mapping (DSM) approaches translates into economic error of management decisions. The error of DSM has previously been assessed by the degree at which an estimated mean prediction deviates from the true mean (Boroughs, 1986), as used in (Behrens et al., 2005), (Nelson et al., 2011) and (Zhu et al., 2015). This error is both a function of the density of soil data and the DSM approach used (Carré et al., 2007). Whilst this measure provides an indication on the accuracy of prediction for a given sample size, further assessment is required to quantify this error economically and to assess the viability of increasing sample size to reduce error. Importantly, linking the prescription of amendment to some expectation of return on investment has been identified as a barrier to the use of such amendments (Bennett and Cattle, 2013; Bennett and Cattle, 2014). Thus, developing the ability to account for the economics of amendment application errors and expected effect provides greater confidence in the soil management decision process.

In considering the accuracy of soil ameliorant recommendations made on the basis of a DSM output, errors of soil ameliorant recommendations can be assessed by quantifying wasted resource, due to over-application, or lost yield potential as a result of under-application. Whilst the cost of over application can be easily quantified, the cost of under-application is more difficult to decipher, as site-specific soil-yield interactions need to be understood to realise the potential yield gap of any given constraint. Previous efforts to quantify crop yield gaps have been focused at the regional or national scale, utilising remotely sensed information with limited soil data (Grassini et al., 2015; Hochman et al., 2013; Neumann et al., 2010; Schierhorn et al., 2014). Even though these approaches provide useful insight at a large spatial scale, they cannot be applied to quantify the site-specific yield gap at the scale required for precision management (i.e. $\approx$12 m agricultural implement frontage). Furthermore, there has been limited attempt to diagnose the relationship between individual soil constraints and the identified yield gap (Dang et al., 2010; Orton et al., 2018), thus providing a similarly limited basis to explore the yield effects of an ameliorant induced constraint change.

Crop-model approaches such as that presented by Dang and Moody (2016), provide the ability to identify the yield gap at a sub-paddock level. These approaches, however, are severely limited by the inability of crop models, such as APSIM (Holzworth et al., 2014), to incorporate the dynamics of soil structural and chemical components (e.g. salinity, sodicity, compaction, acidity) that soil amelioration recommendations are based upon. Therefore, the yield gap

associated with under-application of soil ameliorant cannot directly be assessed using these methods, as it is not possible to observe the yield response of a structurally or chemically induced change directly. A site-specific soil structural yield model is one potential approach to making this assessment.

This chapter aims to investigate the minimum dataset required to economically optimise site-specific agronomic recommendations with an accepted level or error. This will be identified by economically quantifying over-application of soil ameliorants and subsequently aiming to quantify the yield gap associated with under-application, via a site-specific soil structural crop model.

Economic errors of a blanket-rate (BR; homogeneous application of a single rate to a field) approach will be compared against a variable-rate (VR; addressing field areas on an as needed basis) approach to gypsum application. Whilst the BR approach will be based on data obtained using a simulated transect sampling regime at various sampling densities, the VR approach will be based on spatial predictions made using ordinary kriging (OK). Recommendations pertaining to amelioration of soil sodicity for a dryland site in central NSW, Australia will be investigated.

6.2. Methodology

6.2.1. Site description and sampling methods

The investigation site is located within the Warren district of the Macquarie Valley in central NSW, Australia (GR 31°49’40.49” S 148°06’44.56”E). The 108 ha dryland site is managed as a 12 m CTF zero-tillage farming system and is under a winter cropping rotation consisting predominantly of wheat, chickpea and barley. The dominant soil types identified at the site were Kandosol and Dermosol as classified using the Australian Soil Classification System (Isbell). Average annual rainfall for the region is 413 mm.

6.2.2. The soil dataset

The investigation site was grid sampled on a 60x60 m grid, providing 300 grid locations. At each location, a soil core was taken that was subsequently split into 10 cm or 20 cm subsample increments. These samples were analysed at 0–10 cm, 10–20 cm, 20–40 cm and 40–60 cm depth increments for soil structural and chemical components in accordance with Rayment and Lyons (2011). The data was subsequently kriged to a 6x6 m grid using OK as

part of the *automap* package in the R programming environment (Hiemstra and Hiemstra, 2013) to provide a benchmark dataset from which comparisons were made. For the purpose of analysis, it is assumed that the OK intensive benchmark dataset consisting of 300 soil cores (1200 depth based direct data points, and ≈120,000 OK data points) represents the true observed variability at the site. Soil sampling at the site was undertaken in April of 2017.

6.2.3. Spatial prediction methods

OK was adopted as the spatial prediction method for VR agronomic recommendations, with BR recommendations based on a bulked random transect method. Whilst it is noted that random transect sampling is not a spatial prediction method, it is investigated here as the control condition to provide a baseline for economic comparison. This method provides a good representation of the approach most adopted by Australian agricultural enterprises, if soil sampling does occur (Lobry de Bruyn, 2019). These methods were investigated by simulating sampling densities of N=10, 20, 50, 100, 150, 200, 250 and 300 samples for the site. Samples were selected from the benchmark dataset for each simulation. For each sampling density, all methods were simulated a total of 10 times to observe the effects of random parameter initialisation. A brief description of the methods is given below.

6.2.3.A.. Random transect sampling

Random sampling was employed to obtain a paddock average condition from which a spatially uniform recommendation was made (BR). For each simulation, a transect was randomly selected between the north-east and south-west field boundaries. N samples were subsequently selected, equidistant, along the transect, from which average site conditions were obtained and a homogenous, full-field ameliorant rate prescribed (BR).

6.2.3.B. Ordinary kriging

OK was employed to create a continuous map of soil properties. For N samples, the site was randomly stratified in to N strata, from which a sample was randomly selected. OK was then used to fit a variogram to these samples for each soil attribute and interpolate a 6x6 m grid kriged map.

6.2.4. *Gypsum recommendations*

Site-specific gypsum recommendations were made using the widely accepted Oster and Jayawardane (1998) formula, given as:

$$GR = 0.0086.\rho_b.d.CEC.(ESP_i - ESP_j) \quad \text{Equation 6.1}$$

where ρ_b is the bulk density (BD) in Mg/m^3, d is the depth to be treated in m, CEC is soil exchange capacity in mmol$_c$/kg, ESP_i and ESP_j are the observed and target soil exchangeable sodium percentages (ESP). A value of ESP_j=3, as guided by Shainberg et al. (1981), was used at all locations to provide a target benchmark for soil dispersion amelioration at all locations, with a calcium exchange efficiency factor of 75% (Bennett et al., 2016). Gypsum recommendations were calculated for each depth layer to 60 cm.

6.2.5. *Crop model*

The crop model applied to assess the yield gap associated with ameliorant under-application was adopted from Chapter 5. The developed Cubist regression tree model was selected for predictions in this chapter, due to the reasonable generalisation results obtained in Chapter 5, and the reduced likelihood of overfitting, compared to the developed artificial neural network (ANN). The model was trained using 36 soil parameters (i.e. pH, EC, clay %, silt %, sand %, BD, potassium, cation exchange capacity (CEC) and ESP for the 4 depth increments) as variables to predict crop yield (mean of the 2013 and 2015 cropping seasons, removing 2016 as an outlier due to unseasonal wet conditions). The 2016 cropping year was removed due to the above average rainfall received during the cropping year (i.e. 681 mm vs 441 mm average), meaning soil constraint-yield interactions could not be well defined by the model due to water not being a limiting factor. Therefore, for the model to be more representative of cropping years, the 2016 year was removed. Cubist parameters were tuned in the R programming environment using the *Caret* package (Kuhn, 2008; Kuhn et al., 2012). Tuned parameters were 15 committees and 3 neighbours, resulting in an R^2=0.691 for training and R^2=0.691 for validation.

6.2.6. Calculation or recommendation error

The true recommendations for the paddock were assumed as the highest data density of N=300 soil cores, and calculated using the benchmark dataset. Spatial recommendations were subsequently calculated using the OK VR approach and random bulked transect BR approach. The areas and total tonnages of over- and under-application were then identified. Lost yield potential due to under-application was estimated by calculating the difference between the yield potential of the benchmark recommendation and the yield potential of the recommendations made using the VR and BR methods at all sampling densities. This was completed by application of the developed crop model for each 6x6 m pixel at the site.

Application cost of gypsum for the site is presented in Figure 6.2. These figures are representative of the region, as per personal correspondence with growers, advisors and transport carriers. Cost of soil sampling was obtained from the Nutrient Advantage commercial soil analysis laboratory (www.nutrientadvantage.com.au), which is a service utilised by growers and advisors within the immediate district of the investigate site, and is therefore representative of the analysis cost. An analysis cost of $75.00 per sample excluding Australian goods and services tax (GST) was used.

Table 6.1. Estimated cost breakdown of gypsum application for the Warren district of NSW

Bulk price	*Transport*	*Spreading*	*Total*
$33	$62	$15	$110

6.2.7. Soil characteristics

Summary statistics of the 300 directly measured soil cores are presented in Figure 6.3. Soil pH conditions are slightly acidic at the surface and transition to slightly alkaline within the 40–60 cm soil layer. Soil acidity is a likely constraint at the site, however, only in the surface 0–10 cm layer. Mean BD across the site is considered high to severe across all depth layers (Hazelton and Murphy, 2016), with little spatial variance. This signifies a high likelihood of the entire site being compaction constrained. Of the 4 variables presented in Figure 6.3, ESP exhibits the largest spatial variability at the site, with sodicity being described as absent to extreme within each individual depth layer. This suggests that sodicity is a constraint at the site and its presence is highly spatially variable. The mean and standard deviation of the 2-year average wheat yield (2013 and 2015) was 2.04 t/ha and 0.632 t/ha respectively.

Table 6.2. Statistics of measured soil properties at the site for the designated depth increments; Fea., feature; Min, minimum; Max, maximum; Av, average; SD, standard deviation.

Fea.	*0–10 cm*				*10–20 cm*				*20–40 cm*				*40–60 cm*			
	Min	*Max*	*Av*	*SD*	*Min*	*Max*	*Av*	*SD*	*Min*	*Max*	*Av*	*SD*	*Min*	*Max*	*Av*	*SD*
pH	5.27	9.15	6.58	0.64	5.98	9.23	7.52	0.68	6.55	9.45	8.23	0.57	5.98	9.65	8.72	0.56
BD	1.18	1.83	1.47	0.11	1.37	1.84	1.61	0.08	1.01	1.85	1.64	0.08	1.1	1.91	1.68	0.07
CEC	5.78	38.28	16.55	6.39	7.64	39.21	23.31	6.16	10.07	66.5	28.24	5.46	11.05	41.98	29.57	4.53
ESP	0.03	20.86	4.01	3.17	0.13	26.21	5.32	3.75	0.05	30.33	7.36	4.69	0.14	34	10.48	5.9

6.2.8. Economic analysis

The net benefit of gypsum application was calculated as per Equation 6.2. Whilst it is recognised the wheat commodity price fluctuates, an average value of $241/t was kept static across all years. This average wheat commodity price was obtained from ABARES using seasonal averages between 2017 and 2019 for the ASW1 wheat grade (ABARES, 2019). The yield response was also kept static across seasons, as the developed model was not able to detect seasonal difference in yield. This value was inherent to the sampling method and density.

$$Net\ benefit_{i,j,k} = \left(\sum_{i=0}^{t} wheat\ price_i \times yield\ response_{i,j,k}\right) - SC_{j,k} - AM_{j,k} \quad \text{Equation 6.2}$$

where wheat price is the price of wheat in year *i* in $/t, yield response is the paddock total yield response in year *i* based on sampling method and density *j,k* and SC and AM are the total sampling cost and total amendment cost for sampling method and density *j,k*.

Economic analysis for gypsum application was undertaken for surface treatment (i.e. 0–20 cm) and profile treatment (0–60 cm), with the yield response being assessed for both treatment depths. Whilst it is noted that deep placement of gypsum to 60 cm is not currently commercially possible, this should not deter the importance of analysing yield response due to subsurface amelioration. Instead, undertaking this work seeks to further build the business case for commercial equipment by demonstrating potential benefits of the exercise.

6.3. Results

6.3.1. Amendment cost of over application

The VR approach was superior to the BR for full profile treatment of ESP for all sampling densities. This was observed by the error of over- and under-application (Figure 6.1) The financial significance of this error in terms of over-application of gypsum alone was large, with a saving of over >$4,500 at a sampling density of 10, >$17,000 at a density of 20, and >$26,000

at a sampling density of 50 for the VR approach for 0–60 cm treatment (Figure 6.2). Whilst the mean error of the BR approach was less than that of VR at a sampling density of 10 in the surface 0–20 cm layer, the uncertainty due to random initialisation of search parameters at this density overlapped for both methods. This suggests the uncertainty of the methods at sampling densities <10 is too large to conclude that either method is superior, in terms of over-application.

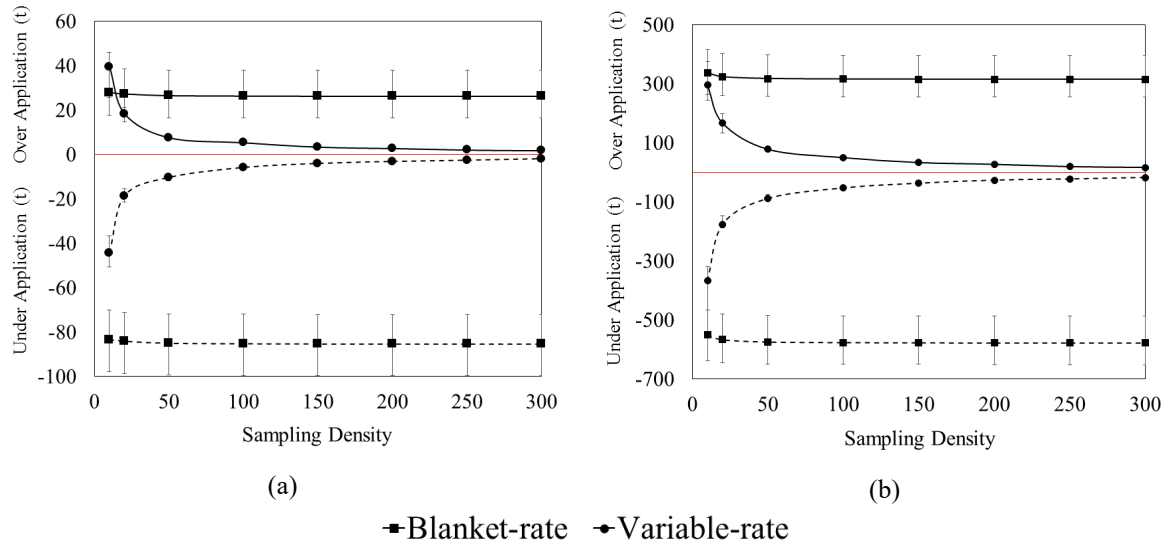


Figure 6.1. Summary of gypsum application based on BR and VR sampling methods for the 0–20 cm topsoil layer (a) and 0–60 cm profile (b)

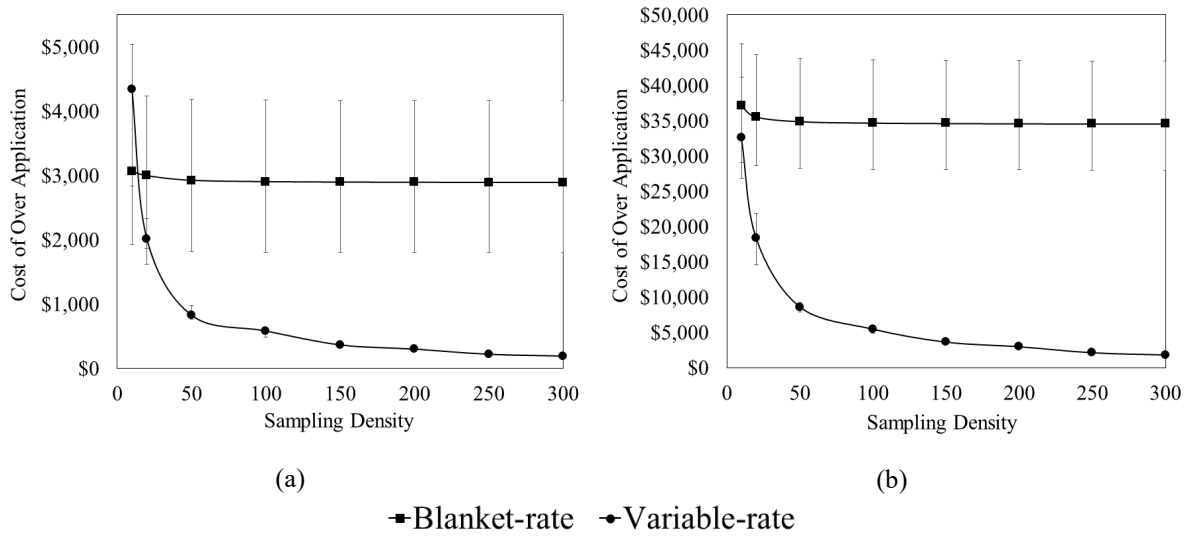


Figure 6.2. Cost of over application of gypsum for the 0–20 cm topsoil layer (a) and 0–60 cm full profile (b). Cost of over application calculated for simulated recommendations based on BR and VR approach for various sampling densities. Error bars represent 1 standard deviation or error between sampling iterations. Price of gypsum taken to be $110/t as suggested by

The BR approach is highly sensitive to random initialisation of the sampling transect, as displayed by the uncertainty of application for both surface (0–20 cm) and profile (0–60 cm) treatments. This uncertainty does not improve with increased sampling, which in is contrast to

the VR approach, where error decreases rapidly to a density of 50 samples. Furthermore, the cost of gypsum over-application does not improve as sampling density increases for BR, which again contrasts that of VR.

6.3.2. Estimating spatial yield response

6.3.2.A. Fitted crop model

Training results of the developed Cubist model are produced in Figure 6.3, with an R^2=0.69 achieved when applied to the entire datasets (i.e. 29,978 observations). Parameters of the model were tuned to 15 committees and 3 neighbours. Generally, the model did not predict well below a yield of 1.5 t/ha, as displayed by the large spread in data below this point. This however only represents a small proportion of the total data, with predictive performance increasing with increasing yield. This suggests that the soil relationships that control yield toward the lower range of yield values are unexplained by the selected variables and relationship.

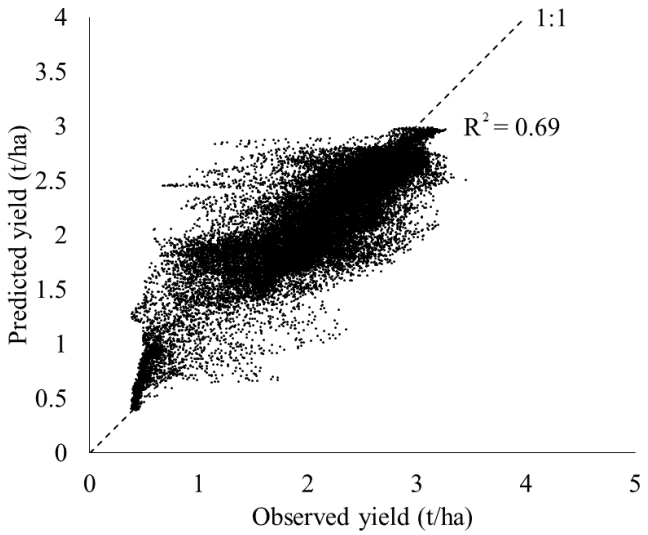


Figure 6.3 Training results of developed Cubist model.

6.3.2.B. Yield response

Gypsum treatment of the soil profile from 0–60 cm resulted in a total simulated yield response much greater than that of treatment constrained to the 0–20 cm topsoil layer (Figure 6.4). This suggests that subsurface sodicity is a larger constraint than surface sodicity, and confirms the severity of higher ESP values measured in the sub-surface layers (Table 6.2).

Whilst a larger response is achieved, the gypsum investment required for amelioration of the 0–60 cm profile is significantly greater than for topsoil treatment alone — total field application of 1,609 t versus 144 t for VR application at N=50 for the full profile and topsoil treatments, respectively.

The simulated yield response due to a VR application is much greater than that achieved by BR at all sampling densities and treatment depths, advocating that the VR approach is far superior to that of BR. If we consider 0–60 cm treatment for both approaches at a sampling density of 50, the magnitude of this difference was 26.2 t of wheat yield per annum, which presents substantial economic difference of $6,414 per annum, in favour of VR, at an average wheat price of $241/t (ABARES, 2019). The mean yield response of the VR application increases with sampling density due to improvements in recommendation accuracy. However, the magnitude of this increase is greatest for full profile treatment between 10 and 50 samples (15.4 t), with only small gains in yield being made by increasing the sample density past 50 samples for the 108 ha site. Additionally, only small differences (3.35 t between 10 and 50 samples) were detected for topsoil only treatment with respect to sampling density.

The spatial yield response — based on 50 samples — due to gypsum treatment is spatially summarised in Figure 6.5 for the 0–20 cm topsoil treatment, and for the 0–60 cm profile treatment in Figure 6.6 . The areas at the site with large under-application errors for the simulated BR application did not provide a large yield increase. In comparison, these same areas provided large yield increases based on the spatially accurate VR application, due to better targeting of the gypsum ameliorant. Interestingly, for the VR treatment, the large majority of yield response was attributed to small regions within the site, whereby treatment of ≈33% of the field accounted for ≈66% of the yield increase (Table 6.3). This highlights the importance of VR approaches in identifying the most economically significant regions for amelioration, which may consist of only small cumulative area, that are highly likely to inadequately and inefficiently treated using BR application.

Importantly, it was observed that the developed yield model predicted a yield decrease due to gypsum application in some areas of the site. For profile treatment, this represented ~3% of the total area. This effect is greatest for the 0–60 cm profile treatment, and was only marginally presented within the full 0–20 cm topsoil.

Upon investigation of these errors, it was found their magnitude was often small (<0.05 t/ha), and only occurred when a small reduction of ESP was induced for a single depth layer,

with the remaining depth layers not requiring treatment. It is therefore assumed that these errors were an artefact of the model, where ESP was not consistently >3 through the profile.

Table 6.3. Percentage of yield attributed to the top performing areas by ha for topsoil and profile amelioration. Yield percentages represent the percent of yield increase achieved by the corresponding best-performing land areas.

Treatment Depth	*Unit*	*Amended field area (ha)*								
		10	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*>80*
Topsoil	% Total yield	25.2	45.37	62.6	76.7	87.25	94.5	97.96	99.7	100
Profile	%Total yield	18.6	34.8	49.56	62.7	73.9	82.9	89.6	94.6	100

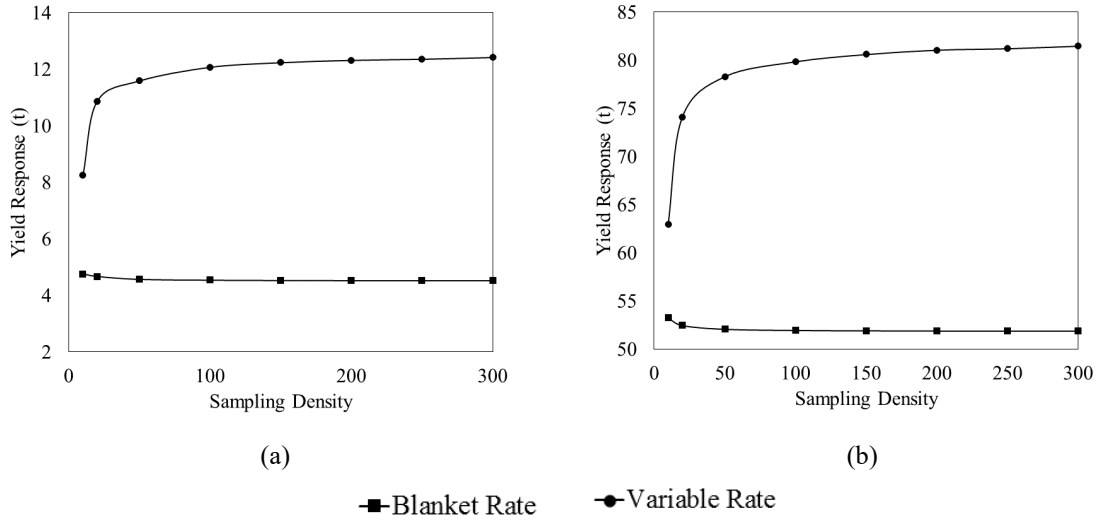


Figure 6.4. Estimated yield response at the investigation site from gypsum amendment application based on transect and kriging spatial prediction methods for the 0–20 cm topsoil layer (a) and 0–60 cm profile (b). Errors bars represent the IQR of 10 iterations of each spatial prediction method. Yield estimated by application of a Cubist regression tree model trained for the investigation site using the average of 2013 and 2014 wheat cropping year

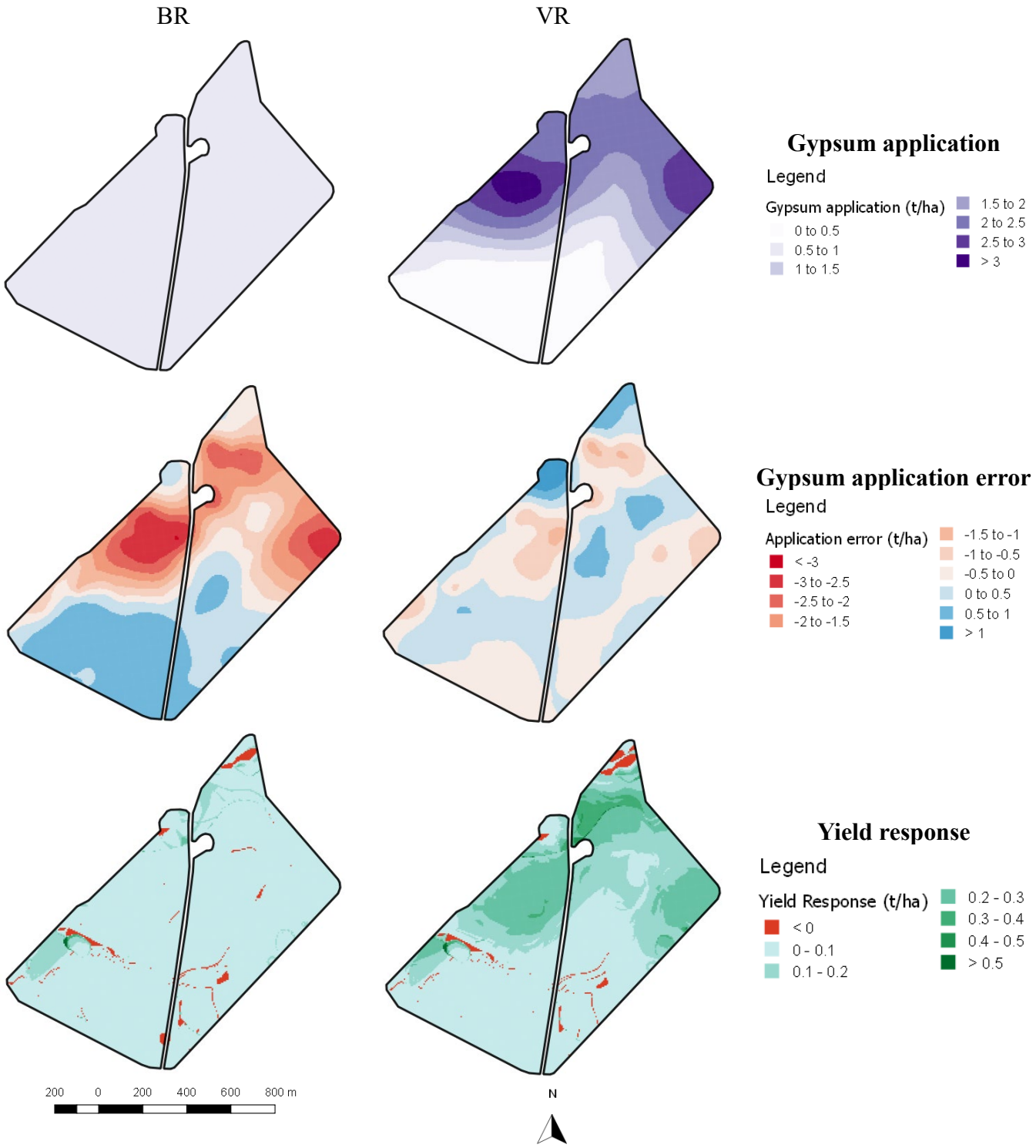


Figure 6.5. Spatial gypsum application, gypsum application error and yield response for 0–20 cm topsoil recommendations based on a BR (left) and variable rate (right) application. Recommendations were based on a sampling density of 20.

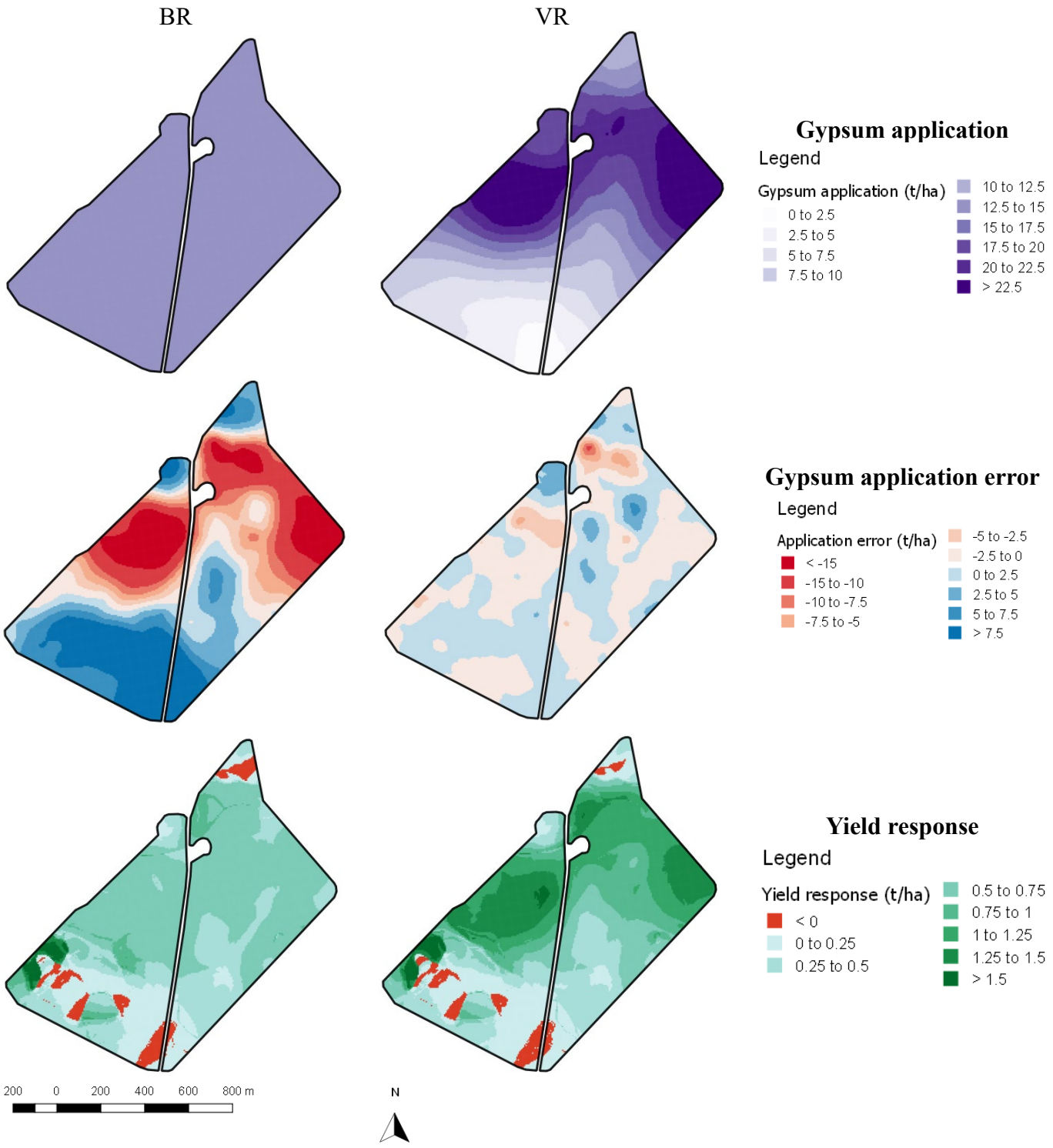


Figure 6.6 Spatial gypsum application, gypsum application error and yield response for 0–60 cm profile recommendations based on a BR (left) and variable rate (right) application. Recommendations were based on a sampling density of 50

6.3.3. Return on investment of VR soil amelioration

Profile treatment to 60 cm achieved a much larger return on investment (ROI) in comparison to treatment limited to 0–20 cm for the simulated gypsum application (Figure 6.7). For profile treatment, the minimum dataset that achieved greatest net profit was 50 core locations (each core consisting of 4 depth sub-samples), which effectively equates to 0.5 cores/ha (1 core/2 ha) for the investigated site. For topsoil treatment, 20 samples was optimal, however, this was only marginally improved over 50. If the amelioration design is focused on the full profile, the results advocates a density of 0.5 cores/ha, while this can be reduced to 0.2 cores/ha where only the topsoil is of interest. Increasing sampling density from 10 to 50 samples using the VR approach equated to an initial saving of >$23,000, due to reduced over-application, as well as a sustained annual yield benefit of 6.68 t of wheat. The difference in net profit for the VR approach between a sampling density of 10 and 50 was >$55,000 after 20 years of payback (0–60 cm), highlighting the importance of increased data collection with economic consideration. Net profit was calculated according to Equation 6.2, where yield response was accumulative over 20 years. At the full profile treatment optimum sampling density of 50 cores (0.5 cores/ha), the rate of return was greatest for the VR approach, as compared to the BR application (Figure 6.8), despite a larger initial investment being required ($160,683 vs $183,401 for the BR and VR, respectively). Furthermore, the time-to-payback for the VR approach was 5 years faster in the topsoil, and 3 years faster for profile treatment, in comparison to BR. After a 20 year time period, the difference in profit between the two approaches was estimated to be >$27,000 and >$104,000 for the 0–20 cm and 0–60 cm treatment layers, respectively. Whilst initially a more expensive option, this highlights the long-term benefits of a variable rate approach to soil amelioration.

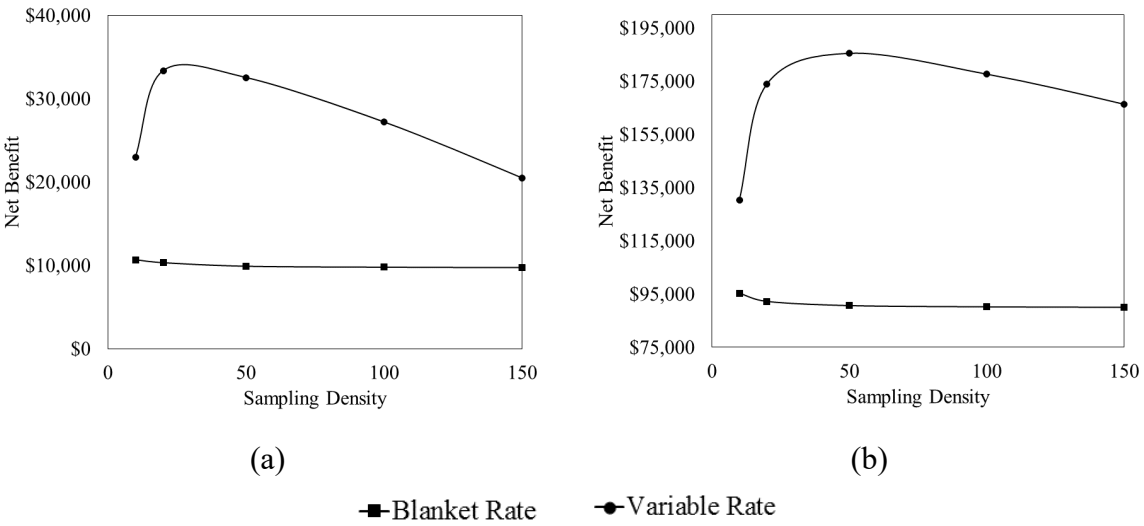


Figure 6.7. Mean net benefit of simulated gypsum application for the 0–20 cm topsoil (a) and 0–60 cm profile (b) layers based on blanket rate and variable rate approaches at varied sampling densities. Net benefit calculated at year 20 after gypsum application and for the mean of 10 simulations of each sampling procedure.

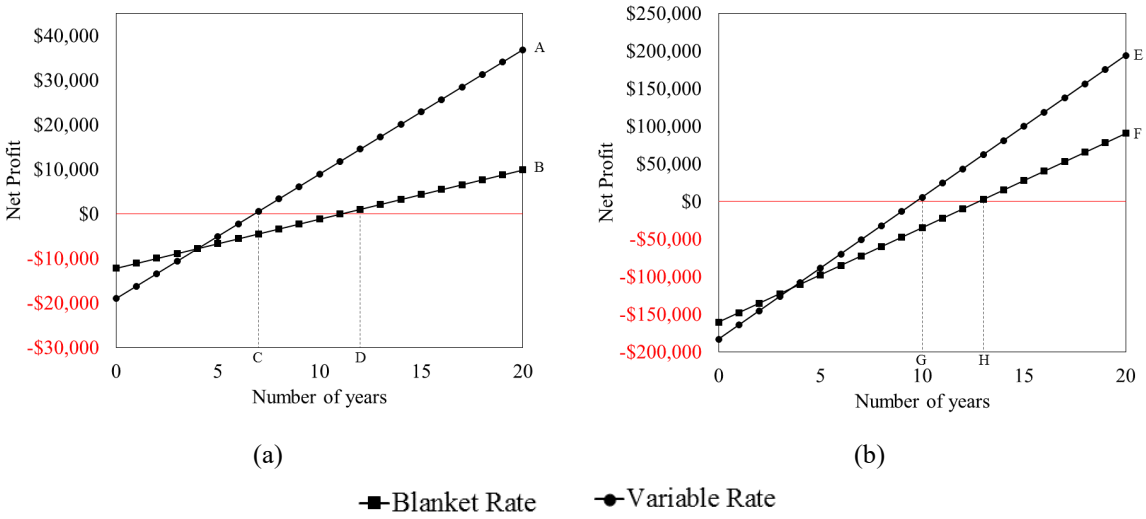


Figure 6.8. Mean ROI for gypsum application over 20 year period for 0–20 cm topsoil (a) and 0–60 cm profile (b) treatment using a blanket rate and variable rate approach. Key points are as follows: A=$36,836, B=$9,841, C=7 years, D=12 years, E=$194,030, F=$90,296, G=10 years, F=13 years. ROI estimated from mean of 10 simulations of each sampling procedure.

6.4. Discussion

6.4.1. The requirement for variable rate application for soil amelioration

In terms of gypsum application to address soil dispersion as a production constraint, a variable rate approach to soil amelioration is far superior to that of blanket rate application. Indeed, while the example used here is gypsum based, this would be true for any constraint management program (Whelan and McBratney, 2000).Whilst over-application using a VR

approach can be reduced, representing an initial saving, the largest benefits are achieved from an increased yield response, which is realised annually. This yield response is the direct cause of a more spatially refined ameliorant application, better targeting the presented soil constraints through more accurate prescription of ameliorant to constraint severity. In the case of dispersive soil amelioration, this subsequently improves soil structure and soil-water dynamics spatially (Bennett et al., 2019; Dang et al., 2018; Rengasamy and Olsson, 1991), therefore increasing crop response (McKenzie et al., 2002a). The economic effect of increased yield response due to a variable rate approach in treating the 0–60 cm profile is substantial, with an increased net return estimated to be >$76,000 after 20 years, as compared to a blanket rate application (i.e. $221,500 vs $145,000 for VR and BR, respectively). However, the operational costs of ameliorant application at subsoil depth need to be better accounted for in future work. Additionally, the model assumes instant benefit due to the change of ESP, which would be more gradual in actuality. Even with these considerations, there is clearly a value proposition in the amelioration of soil profiles when undertaking a VR approach to soil amelioration.

Importantly, it was observed here that the majority of yield increases due to amelioration are achieved from small areas within a field. Soil constraints may vary significantly at small spatial scales (Dang and Moody, 2016; McBratney and Pringle, 1999), meaning the severity of a constraint can be spatially detailed at a scale much finer than investigated for this single field; i.e. the resolution of variable rate will need to be determined by the limitation of the equipment to apply it, rather than the inherent variability of the soil. Additionally, the scale of the farming system, accompanied with the net income per area, will determine the ability to increase the sampling density to better understand this more detailed variability; e.g. for smaller scale high-value horticulture it may be realistic to manage on a square meter basis as opposed to a broad-acre enterprise and a hectare basis. Furthermore, given the nonlinear relationship between soil constraints and yield (Dai et al., 2011; Drummond et al., 1998; Sudduth et al., 1996), an equivalent change in constraint between severe and less severe regions may not result in an equivalent yield response. Consequently, small areas within a site can be highly economically significant, in terms of their yield potential, as shown here. Identifying these regions subsequently should become imperative for improved management. Therefore, a requirement exists to employ spatial sampling strategies that can accurately detect specific sub regions of high economic significance, at the specific resolution that is both economically feasible and scalable to the agricultural context.

The importance of a variable rate approach for soil amelioration is increased with depth, with the greatest economic return achieved for VR where full profile treatment (0–60 cm) occurred, as opposed to topsoil treatment (0–20 cm). The influence of subsoil constraints on crop performance can be immense (Orton et al., 2018; Rengasamy, 2002), and the cost of treatment is often substantial, due to both an increased ameliorant requirement and logistics of sub-surface treatment. Consequently, small errors in subsurface recommendations can be economically substantial. Therefore, it is imperative to accurately characterise subsurface ameliorant recommendations and applications to maintain economic efficacy of profile treatment. This requires sub-surface data collection, which must remain a focus of future work. Whilst it is recognised here that the deep placement of gypsum products is not widely practiced, the cause of this is largely due to the absence of commercial deep application equipment, minimal investigation of accurate rate application with depth (Jayawardane and Blackwell, 1985; Jayawardane et al., 1988) and uncertainties pertaining to the return on investment.

The amelioration of dispersive soil via the application of gypsum has been investigated here. However, the same approach can be applied to other constraints and inputs, such as acidity, compaction, salinity and nutrient deficiencies/toxicities. Whilst not adopted currently, opportunity exists for VR strategic deep tillage to remediate subsoil compaction, whereby the depth to ripping is adjusted with consideration toward depth of compaction and presence of hostile subsoils. This approach could also be coupled to apply gypsum and/or lime, or indeed any other blend of input, within the same pass given the VR technology currently available; it is simply a matter of engineering development and the simultaneous development of an existing market to adopt this.

6.4.2. Optimising soil sampling investment

For the site investigated, the optimum number of soil sample locations for the treatment of sodicity to 60 cm was considered to be 50, effectively equivalent to 0.5 soil cores/ha (1 core/2 ha), with assessment of four depth layers (4 soils samples/2 ha). At an analysis cost of $75/sample (Nutrient Advantage®, 2019), this equated to an investment of $139/ha for the 108 ha site. Whilst this investment may seem substantial given current perceptions surrounding sampling investment, the annual yield benefit was 26.2 t per annum greater than a BR application, which only required an initial sampling investment of $2.80/ha, thus highlighting the opportunity cost of data limiting approaches to agronomic recommendations.

The optimal sampling density identified here has previously been considered insufficient for spatial prediction of soil properties using OK (Webster and Oliver, 1992), on which the presented recommendations were based on. This is due to the noise at low sampling densities which was attributed to large error when fitting the variogram, therefore, contributing error to spatial predictions (Webster and Oliver, 1992). However, in the context presented here, which is cost limiting, increasing sampling above 20–50 locations is cost prohibitive, as the cost of data acquisition exceeds that of the benefit, due to reduced spatial prediction error. Additionally, the area over which the samples are taken versus the inherent variability of that area becomes important; i.e. if 20–50 samples were taken over 100 square kilometres, as compared to 100 ha, with the aim of providing spatial ameliorant application rates, then the accuracy and usefulness of these two maps would differ substantially. This means that the sampling density should be considered on the basis of the variability, context of the problem and economic feasibility of increasing data further.

Demonstrated here is the requirement for increased soil data collection in the x,y and z planes in agriculture, which has traditionally been considered less important due to its perceived usefulness (Bennett, 2015). However, taking a data limiting approach to soil amelioration recommendations can be economically detrimental, the effects of which are amplified over time. Using examples similar to that presented in this body of work, there is a requirement to change the way data is valued in Australian agriculture (Bennett and Cattle, 2013; Bennett and Cattle, 2014; Lobry de Bruyn, 2019) in order to drive on-farm efficiency gains through improved variable rate management. Subsequently, realising this value provides opportunity to reduce the costs of sample analysis, due to an increased demand to drive laboratory throughput efficiencies to match that of other global agricultural industries. This in-turn will allow for optimum sampling densities to be re-calibrated to drive the accuracy of on-farm soil recommendations further.

The minimum sampling requirement for agronomic recommendations is site-specific, and is dependent on the accuracy of spatial predictions, the cost of sample collection and analysis, the cost of soil amendment and application, depth of treatment and the yield response due to amelioration. Whilst the latter factors may be estimated and accounted for irrespective of site conditions, the accuracy of spatial predictions is directly influenced by the inherent variability presented at the site. With this in mind, I hypothesize the development of a metric that describes the spatial variability of a new site in an attempt to inform optimal sampling

requirements with minimal investment. This metric will likely combine spatial data streams such a yield data, satellite data and proximally sensed information in a hybrid approach to recommend optimised sampling schemes with economic consideration toward the context of application.

6.4.3. Limitations of and opportunities for machine learning to predict yield in response to constraint amelioration

A recognised limitation of the presented yield prediction model is the inability to assess the uncertainty of yield response as a function of spatiotemporal variability (e.g. climate), which is known to be significant (Mcbratney et al., 2007). Whilst an average of two wheat cropping years were used for training, there is no guarantee that these appropriately represent seasonal conditions at the site, and instead, the model will be biased to these years. Building a more robust model requires increased yield information to explore the interactions of weather on soil-crop dynamics (concluded in the previous chapter). This would allow for improved financial forecasting when assessing the uncertainty pertaining to the ROI of soil amelioration. While this improvement certainly must be made, the current results demonstrate that such an approach would be achievable and useful, providing a means to develop a data collection and ameliorant application regime value proposition in the meantime.

Interestingly, the developed model predicted a yield decline at ~3% of the site by area, following gypsum application. Given the extensive literature on the use of gypsum application to dispersive soil (Davidson and Quirk, 1961; Kazman et al., 1983; Oster, 1982; Shainberg et al., 1982), and the effect of calcium on soil structure and plant function (Aylmore et al., 1971; White and Broadley, 2003) this result not consistent and doesn't align with expectations. Whilst it is possible to build non-negative logic into cubist models to ensure negative values are not predicted, identifying the samples causing the negative results can be pertinent to further explore relationships associated with these soil samples (Minasny and McBratney, 2008). As the negative results are spatially correlated, further investigation may aid understanding of unique mechanistic relationships, or if this is purely the result of error within the data. If true, the former could highlight the interaction complexity between soil constraints (Bennett et al., 2015a; Dang et al., 2008; Lawes et al., 2009; Nuttall and Armstrong, 2010; Nuttall et al., 2003), where mechanistically, a change in one constraint may influence a subsequent change in another (e.g. increasing EC to combat ESP). It is therefore advantageous to explore the interactions of constraints with yield simultaneously (Dang et al., 2010; Orton et al., 2018), and

signifies a requirement for soil science domain knowledge in interpreting the model outputs; an completely autonomous approach to data interpretation would be fraught with error and practical danger for soil degradation at this point in time.

Assuming the developed model is correct, the interaction-effects of soil constraints on crop yield can be investigated. Whilst we have adjusted ESP individually to observe the yield effects, opportunity exists to employ a search algorithm to the model to identify the soil parameters that maximize yield at each individual location within the site. Furthermore, the integration of an economic model will allows for a constrained optimization problem whereby yield is maximized with consideration toward the cost of amelioration, therefore providing a means for optimised amendment recommendation. However, that was beyond the scope of the immediate work.

Achieving the above requires the assumption of generalisation to be satisfied, whereby the model can accurately predict the yield response due to the removal or alleviation of a constraint. In order to accomplish this, model training requires sufficient examples that occupy unconstrained conditions to learn from. Without these data, the model is forced into predicting beyond the bounds of training when simulating the amelioration of a constraint, which is known is problematic (Kuhn and Johnson, 2013) and is strongly advised against. In pursuing this, one must consider the likelihood of achieving unconstrained conditions, particularly at the field scale, as in reality, these may be non-existent (e.g. the lack of initial benchmark for soil compaction; Antille et al., 2016a; Bennett et al., 2015b). In this situation, we are constrained by data availability, not in terms of volume, but variability within the observations. An example of this is presented for soil compaction within the investigated site presented here. Considering the thresholds of BD to infer compaction severity within clay soils (Hazelton and Murphy, 2016), the mean values at all depths generally suggested high to severe compaction is present. Soil BD naturally increases with depth — due to the mass of over-burden soil and gravity — and separating this natural phenomena from machine induced compaction is no easy task. Furthermore, in comparison to other features (e.g. ESP), BD does not exhibit great spatial variation within either of the depths. This suggests the large majority of the site is likely to be compaction constrained, due to years of random traffic farming prior to implementation of controlled traffic farming without soil renovation (Bennett et al., 2017; McGarry et al., 1999; Tullberg et al., 2007). Therefore, limited unconstrained observations have the potential exist

for a number of constraints, and can affect some constraints more than others, subsequently inhibiting a model's ability to learn the response due to constraint amelioration.

6.5. Conclusion

The work presented here aimed to identify the minimum soil sampling requirements to optimise soil amelioration with economic consideration, using the treatment of dispersive soil as an example. For the investigated site, a sampling density of 20 and 50 core locations using a VR approach resulted in the economically optimal density for treatment to 20 and 60 cm, respectively. This translated to an investment of ≈$28 and ≈$138 /ha. Whilst this investment is substantial, treatment to 60 cm achieved a net positive gain of ~$104,000 using VR, as compared to the BR approach, over a 20 year period, with the payback period of amelioration being realised at 10 years (including the cost of sampling and ameliorant), 3 years faster than BR. Furthermore, it was identified that small areas contribute comparatively large yield responses and are of high economic significance. Hence, a VR approach to soil amelioration is paramount.

A site-specific yield model based on the Cubist model-tree approach was developed, predicting a spatial yield response to simulated gypsum application. The model provides reasonable predictions to make investment decisions around obtaining soil data information at a spatial scale and the ability to employ recommendations based on these with some expectation of return on investment. However, the model was limited to using 2 years of yield information as training examples, meaning the uncertainty of yield response due to climatic variability cannot be assessed. There remains a requirement to build in additional yield information to learn the spatiotemporal variations for this model, as well as for the development of future models at other sites.

Most importantly, with the observed benefits of increased soil data collection, a requirement exists within industry to change the thinking of how soil data investment is perceived. This work provides an example of how the value proposition can be built to support claims that greater data density and spatial understanding can result in localised yield benefits overtime. Whilst a number of assumptions were made in during this work (i.e. Cubist model predictions, kriging predictions, representative rainfall, gypsum dissolution efficiency etc.), these were necessary as an initial step to demonstrate the ROI of soil sample investment. Future work should consider the sensitivity of ROI to factors such as rainfall, response of varied crop types and dissolution rate. As was demonstrated in this work, soil sampling and data analysis

must be considered a capital investment with strategic, long-term ROI plans, rather than simply an operational cost. Changing this mind-set is imperative for the progression of precision agriculture.

6.6. References

ABARES, 2019. Weekly Australian Climate, Water and Agricultural Update. Department of Agriculture and Water Resouces.

Abbott, T., McKenzie, D., 1986. Improving soil structure with gypsum. Department of Agriculture. New South Wales, Agfact AC 10.

Antille, D.L., Bennett, J.M., Jensen, T.A., 2016. Soil compaction and controlled traffic considerations in Australian cotton-farming systems. Crop and Pasture Science 67(1), 1-28.

Aylmore, L., Karim, M., Quirk, J., 1971. Dissolution of gypsum, monocalcium phosphate, and superphosphate fertilizers in relation to particle size and porous structure. Soil Research 9(1), 21-32.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Journal of plant nutrition and soil science 168(1), 21-33.

Bennett, J.M., 2015. Agricultural big data: utilisation to discover the unknown and instigate practice change. Farm Policy Journal 12(1), 43-50.

Bennett, J.M., Cattle, S., 2013. Adoption of soil health improvement strategies by Australian farmers: I. Attitudes, management and extension implications. The Journal of Agricultural Education and Extension 19(4), 407-426.

Bennett, J.M., Cattle, S., 2014. Adoption of soil health improvement strategies by Australian farmers: II. Impediments and incentives. The Journal of Agricultural Education and Extension 20(1), 107-131.

Bennett, J.M., Cattle, S., Singh, B., 2015a. The efficacy of lime, gypsum and their combination to ameliorate sodicity in irrigated cropping soils in the Lachlan Valley of New South Wales. Arid Land Research and Management 29(1), 17-40.

Bennett, J.M., Marchuk, A., Raine, S., Dalzell, S., Macfarlane, D., 2016. Managing land application of coal seam water: A field study of land amendment irrigation using saline-sodic and alkaline water on a Red Vertisol. Journal of environmental management 184, 178-185.

Bennett, J.M., Roberton, S., Marchuk, S., Woodhouse, N., Antille, D., Jensen, T., Keller, T., 2019. The soil structural cost of traffic from heavy machinery in Vertisols. Soil and Tillage Research 185, 85-93.

Bennett, J.M., Roberton, S.D., Jensen, T.A., Antille, D.L., Hall, J., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: I. Heavy machinery impact on the soil resource. Soil and Tillage Research 168, 143-154.

Bennett, J.M., Woodhouse, N.P., Keller, T., Jensen, T.A., Antille, D.L., 2015b. Advances in cotton harvesting technology: a review and implications for the John Deere round baler cotton picker. Journal of Cotton Science 19(2), 225-249.

Boroughs, P., 1986. Principles of Geographic Information Systems for Land Resource Assessment. Beckett, PHT.

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007. Digital soil assessments: Beyond DSM. Geoderma 142(1-2), 69-79.

Dai, X., Huo, Z., Wang, H., 2011. Simulation for response of crop yield to soil moisture and salinity with artificial neural network. Field Crops Research 121(3), 441-449.

Dang, A., Bennett, J.M., Marchuk, A., Biggs, A., Raine, S., 2018. Quantifying the aggregation-dispersion boundary condition in terms of saturated hydraulic conductivity reduction and the threshold electrolyte concentration. Agricultural water management 203, 172-178.

Dang, Y., Dalal, R., Mayer, D., McDonald, M., Routley, R., Schwenke, G., Buck, S., Daniells, I., Singh, D., Manning, W., 2008. High subsoil chloride concentrations reduce soil water extraction and crop yield on Vertosols in north-eastern Australia. Australian Journal of Agricultural Research 59(4), 321-330.

Dang, Y., Dalal, R., Routley, R., Schwenke, G., Daniells, I., 2006. Subsoil constraints to grain production in the cropping soils of the north-eastern region of Australia: an overview. Australian Journal of Experimental Agriculture 46(1), 19-35.

Dang, Y., Moody, P., 2016. Quantifying the costs of soil constraints to Australian agriculture: a case study of wheat in north-eastern Australia. Soil Research 54(6), 700-707.

Dang, Y.P., Dalal, R.C., Buck, S., Harms, B., Kelly, R., Hochman, Z., Schwenke, G.D., Biggs, A., Ferguson, N., Norrish, S., 2010. Diagnosis, extent, impacts, and management of subsoil constraints in the northern grains cropping region of Australia. Soil Research 48(2), 105-119.

Davidson, J., Quirk, J., 1961. The influence of dissolved gypsum on pasture establishment on irrigated sodic clays. Australian Journal of Agricultural Research 12(1), 100-110.

Doyle, R., Habraken, F., 1993. The distribution of sodic soils in Tasmania. Soil Research 31(6), 931-947.

Drummond, S., Joshi, A., Sudduth, K.A., 1998. Application of neural networks: precision farming, Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on. IEEE, pp. 211-215.

Ford, G., Martin, J., Rengasamy, P., Boucher, S., Ellington, A., 1993. Soil sodicity in Victoria. Soil Research 31(6), 869-909.

Grassini, P., van Bussel, L.G., Van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., de Groot, H., van Ittersum, M.K., Cassman, K.G., 2015. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. Field Crops Research 177, 49-63.

Greene, R., Ford, G., 1985. The effect of gypsum on cation exchange in two red duplex soils. Soil Research 23(1), 61-74.

Hazelton, P., Murphy, B., 2016. Interpreting soil test results: What do all the numbers mean? CSIRO publishing.

Hiemstra, P., Hiemstra, M.P., 2013. Package 'automap'. compare 105, 10.

Hochman, Z., Gobbett, D., Holzworth, D., McClelland, T., van Rees, H., Marinoni, O., Garcia, J.N., Horan, H., 2013. Reprint of "Quantifying yield gaps in rainfed cropping systems: A case study of wheat in Australia". Field Crops Research 143, 65-75.

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM – Evolution towards a new generation of agricultural systems simulation. Environmental Modelling & Software 62, 327-350.

Jayawardane, N., Blackwell, J., 1985. The effects of gypsum-enriched slots on moisture movement and aeration in an irrigated swelling clay. Soil Research 23(4), 481-492.

Jayawardane, N., Blackwell, J., Blackwell, P., 1988. Fragment size distribution within slots created in a duplex soil by a prototype rotary slotter. Soil and Tillage Research 12(1), 53-64.

Kazman, Z., Shainberg, I., Gal, M., 1983. Effect of low levels of exchangeable sodium and applied phosphogypsum on the infiltration rate of various soils 1. Soil Science 135(3), 184-192.

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

Kuhn, M., Johnson, K., 2013. Applied predictive modeling, 26. Springer.

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

Lawes, R., Oliver, Y., Robertson, M., 2009. Capturing the in-field spatial–temporal dynamic of yield variation. Crop and Pasture Science 60(9), 834-843.

Lobry de Bruyn, L., 2019. Learning opportunities: Understanding farmers' soil testing practice through workshop activities to improve extension support for soil health management. Soil Use and Management.

McBratney, A., Pringle, M., 1999. Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. Precision Agriculture 1(2), 125-152.

Mcbratney, A.X., Whelan, B.M., Shatar, T.M., 2007. Variability and uncertainty in spatial, temporal and spatiotemporal crop‐yield and related data, Ciba Foundation Symposium 210‐Precision Agriculture: Spatial and Temporal Variability of Environmental Quality: Precision Agriculture: Spatial and Temporal Variability of Environmental Quality: Ciba Foundation Symposium 210. Wiley Online Library, pp. 141-160.

McGarry, D., Sharp, G., Bray, S., 1999. The current status of soil degradation in Queensland cropping soils. Report No. DNRQ990092. Queensland Department of Natural Resources, Brisbane, Qld.

McKenzie, D., Abbott, T., Chan, K., Slavich, P., Hall, D., 1993. The nature, distribution and management of sodic soils in New-South-Wales. Soil Research 31(6), 839-868.

McKenzie, D., Bernardi, A., Chan, K., Nicol, H., Banks, L., Rose, K., 2002. Sodicity v. yield decline functions for a Vertisol (Grey Vertosol) under border check and raised bed irrigation. Australian Journal of Experimental Agriculture 42(3), 363-368.

Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemometrics and Intelligent Laboratory Systems 94(1), 72-79.

Naidu, R., Merry, R.H., Churchman, G., Wright, M., Murray, R., Fitzpatrick, R.W., Zarcinas, B., 1993. Sodicity in South Australia-a review. Soil Research 31(6), 911-929.

Nelson, M., Bishop, T., Triantafilis, J., Odeh, I., 2011. An error budget for different sources of error in digital soil mapping. European Journal of Soil Science 62(3), 417-430.

Neumann, K., Verburg, P.H., Stehfest, E., Müller, C., 2010. The yield gap of global grain production: A spatial analysis. Agricultural systems 103(5), 316-326.

Nuttall, J., Armstrong, R., 2010. Impact of subsoil physicochemical constraints on crops grown in the Wimmera and Mallee is reduced during dry seasonal conditions. Soil Research 48(2), 125-139.

Nuttall, J.G., Armstrong, R., Connor, D., 2003. Evaluating physicochemical constraints of Calcarosols on wheat yield in the Victorian southern Mallee. Australian Journal of Agricultural Research 54(5), 487-497.

Orton, T.G., Mallawaarachchi, T., Pringle, M.J., Menzies, N.W., Dalal, R.C., Kopittke, P.M., Searle, R., Hochman, Z., Dang, Y.P., 2018. Quantifying the economic impact of soil constraints on Australian agriculture: A case‐study of wheat. Land Degradation & Development 29(11), 3866-3875.

Oster, J., 1982. Gypsum usage in irrigated agriculture: a review. Fertilizer research 3(1), 73-89.

Oster, J., Jayawardane, N., 1998. Agricultural management of sodic soils.

Rayment, G.E., Lyons, D.J., 2011. Soil chemical methods: Australasia, 3. CSIRO publishing.

Rengasamy, P., 2002. Transient salinity and subsoil constraints to dryland farming in Australian sodic soils: an overview. Australian Journal of Experimental Agriculture 42(3), 351-361.

Rengasamy, P., Olsson, K., 1991. Sodicity and soil structure. Soil Research 29(6), 935-952.

Schierhorn, F., Faramarzi, M., Prishchepov, A.V., Koch, F.J., Müller, D., 2014. Quantifying yield gaps in wheat production in Russia. Environmental Research Letters 9(8), 084017.

Shainberg, I., Keren, R., Frenkel, H., 1982. Response of Sodic Soils to Gypsum and Calcium Chloride Application 1. Soil Science Society of America Journal 46(1), 113-117.

Shainberg, I., Rhoades, J., Prather, R., 1981. Effect of Low Electrolyte Concentration on Clay Dispersion and Hydraulic Conductivity of a Sodic Soil 1. Soil Science Society of America Journal 45(2), 273-277.

Shaw, R., Brebber, L., Ahern, C., Weinand, M., 1994. A review of sodicity and sodic soil behavior in Queensland. Soil Research 32(2), 143-172.

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

Tennant, D., Scholz, G., Dixon, J., Purdie, B., 1992. Physical and chemical characteristics of duplex soils and their distribution in the south-west of Western Australia. Australian Journal of Experimental Agriculture 32(7), 827-843.

Tullberg, J., Yule, D., McGarry, D., 2007. Controlled traffic farming—from research to adoption in Australia. Soil and Tillage Research 97(2), 272-281.

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

Whelan, B., McBratney, A., 2000. The “null hypothesis” of precision agriculture management. Precision Agriculture 2(3), 265-279.

White, P.J., Broadley, M.R., 2003. Calcium in plants. Annals of botany 92(4), 487-511.

Zhu, A., Liu, J., Du, F., Zhang, S., Qin, C., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. European Journal of Soil Science 66(3), 535-547.

7. A Bayesian approach toward the use of qualitative information to inform on-farm decision making: The example of soil compaction

7.1 Introduction

Situations often arise when the volume and variety of data is insufficient to explore constraint-yield interactions using empirical methods. Opportunity exists to augment the available limited site-specific data, with existing knowledge of the relationships, to provide localised inference for constraint management. The inclusion of management practice nuance on yield production requires vast collection of data at the full range of variability in order for it to be built into quantitative linear and non-linear empirical models. However, the situation is such that by the time sufficient management nuance information had been collected, if that is even possible, the value of the output is now largely redundant in terms of the ability to mitigate impacts. A good example of this is soil compaction, which is a construct, rather than a direct soil characteristic. To understand the management effects of machinery on soil compaction spatially a benchmark condition is required (initial uncompacted state) and then the load must be imposed to measure change in density; i.e. the soil damage is done in the collection of data. Furthermore, there is normally no know benchmark condition due to this information having not been collected prior to agricultural inception (Antille et al., 2016a), meaning the best that can be achieved is some estimate of the likelihood of compaction given management history. Due to the qualitative and semi-quantitative nature of this knowledge, a hybrid approach that is capable of integrating these data sources is required. Bayesian Belief Networks (BBN) offer a probabilistic approach to merging such data streams (Smith et al., 2007a). Both predictive and diagnostic reasoning can be performed to explore the interaction between variables within a system with the aim of identifying management critical control points (Henriksen and Barlebo, 2008; Robertson and Wang, 2004). Therefore, this chapter aims to adopt a BBN approach in exploring constraint-yield interactions by merging existing knowledge with quantitative information to provide local inference tin terms of soil compaction risk.

Soil compaction is a major limiting constraint in conventional agriculture (Hamza and Anderson, 2005; Orton et al., 2018). It is difficult to quantify soil compaction using quantitative metrics (e.g. bulk density [BD] and penetration resistance). This requires additional knowledge of site characteristics that are rarely available, such as clay mineralogy and some reference to an un-compacted state in order to benchmark its severity, as well as a site-specific understanding of the site-specific soil moisture interactions which effect resistance

and shrink-swell properties. Furthermore, due to the complex interactions of soil constraints (Bennett et al., 2015a; Dang et al., 2008; Lawes et al., 2009; Nuttall and Armstrong, 2010; Nuttall et al., 2003), it is often difficult to quantify the influence of compaction on yield without detailed field trials (Bartimote et al., 2017). Hence, the effects of compaction are often overlooked in the management of farming systems with field operations frequently undertaken at less than ideal moisture conditions (Bennett et al., 2019; Braunack and Johnston, 2014) with machines often exceeding >30 Mg (Bennett et al., 2017).

Whilst controlled-traffic farming (CTF) provides best management practice (BMP) for compaction management its uptake is relatively limited (Tullberg et al., 2007) due to its perceived level of importance. For soils with high clay content, Roberton and Bennett (2017)and Bennett et al. (2019) demonstrated that there is effectively no safe soil traffic window for harvesting using heavy modern day broad-acre machinery with axle loads exceeding 200 kPa (Håkansson, 1990). Safe traffic requires the permanent wilting point to have been reached throughout the full profile (Bennett et al., 2019) (Braunack and Johnston, 2014) suggest this is highly unlikely for irrigated agriculture and is achievable <50% of the time in dryland agriculture. Therefore, it is advantageous to develop means to assess the likely yield consequences, based upon compaction risk and severity in a random traffic farming system (RTF). Thus, highlighting the requirement for CTF in conventional agriculture..

A number of empirical and mechanistic methods exist that predict soil deformation under varied loading conditions (Bailey et al., 1995; Gupta and Larson, 1982; O'sullivan et al., 1999). Whilst these models provide reasonable predictions of soil deformation in European soils (Keller et al., 2007), their applicability for Australian soils is impeded due to the lack of localised empirical training data. Furthermore, they provide limited ability to quantify the resulting BD in terms of its severity and risk of system impact within the context of agriculture (e.g. yield effects of increased soil density). Recognising the limitations of empirical approaches for environmental modelling, Troldborg et al. (2013) developed a generic framework for risk-based assessments in data limited systems. The authors applied a BBN approach to predict soil compaction risk using climate, soil and management information. The model categorically predicts the risk of compaction at a national scale for the whole of Scotland, using broad assumptions surrounding site, weather and management conditions to parameterise the model. An opportunity exists however to supplement these assumptions with existing biophysical models that parameterise soil loading and soil moisture components of the network.

This will allow for more localized soil compaction inference. Furthermore, it is prudent to associate the compaction risk output with a yield decline metric to better realise the impacts of compaction in an effort to demonstrate the value proposition of system change to enable compaction mitigation (i.e. expenditure to convert to a CTF system). The aim of this paper is to adopt a BBN approach in exploring constraint-yield interactions by merging existing knowledge with quantitative information to provide field scale inference towards the of soil compaction risk under various loading and traffic conditions. . A hybrid modelling approach incorporating the PERFECT water balance model (Littleboy et al., 1989)and the SoilFlex stress-state distribution model (Keller et al., 2007) is used to help locally parameterise the model. The developed network will be employed to assess the risk of soil compaction and associated yield declines across various locations in Northern NSW, Australia. This will allow for the semi-quantitative risk assessment of soil compaction. The model will be used to assess the risk of soil compaction and associated yield declines across various locations in Northern NSW, Australia.

7.2 Methodology

This methodology section describes the development of a BBN to estimate soil compaction risk and the associated impacts in terms of crop yield decline. To demonstrate the model, simulations will be completed at across the dominant cropping regions of New South Whales in Australia (Figure 7.1)

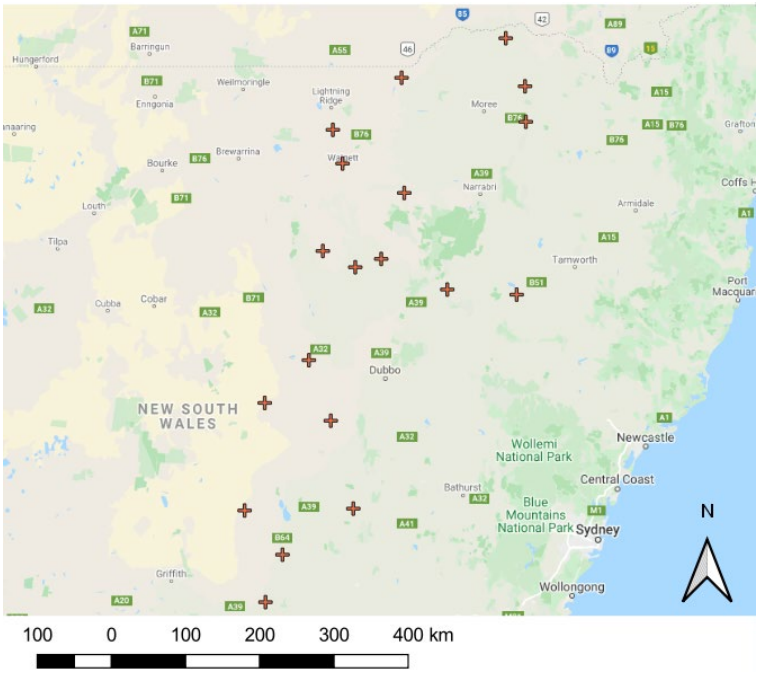


Figure 7.1 Locations where BBN model was simulated for compaction risk and assoicated yield decline.

7.2.1 Model Development

The structure of the applied BBN model was modified from the work of Troldborg et al. (2013) to include existing empirical and biophysical models for parameterisation of the *Total exposure* and *Soil wetness* nodes of the network (Figure 7.2). This allowed for compaction risk predictions at the field-scale, using directly measured information (e.g. weather data). Total exposure was inferred by applying the SoilFlex stress distribution model (Keller et al., 2007). This requires machinery loading characteristics to be supplied including wheel load (kg), tyre inflation pressure (kPa), tyre width (cm) and tyre diameter (m). The output of the SoilFlex is stress estimates (kPa) at 5 cm increments to a depth of 150 cm. To parameterize the Total Exposure node of the BBN the soil stress estimates (kPa) were categorised into 'Low', 'Medium' and 'High' bins using thresholds reported by Horn and Fleige (2003), . The stress ranges for each category are presented in Table 7.1. The *Soil wetness* node of the BBN was parameterized by application of APSIM (Keating et al., 2003b) which adopts the PERFECT water balance model (Littleboy et al., 1989) to estimate soil moisture conditions.. The categorisation of soil moisture (Table 7.1) was based on published data and expert opinion (Bennett et al., 2019).

Table 7.1 Boundaries of total exposure categories based on soil stress and the soil wetness categories based on gravimetric soil moisture content

Total exposure		*Soil wetness*	
Category	*Stress range (kPa)*	*Category*	*Moisture content (cm^3/cm^3)*
Low	<60	Low	<20
Medium	60–120	Medium	20–30
High	>120	High	>30

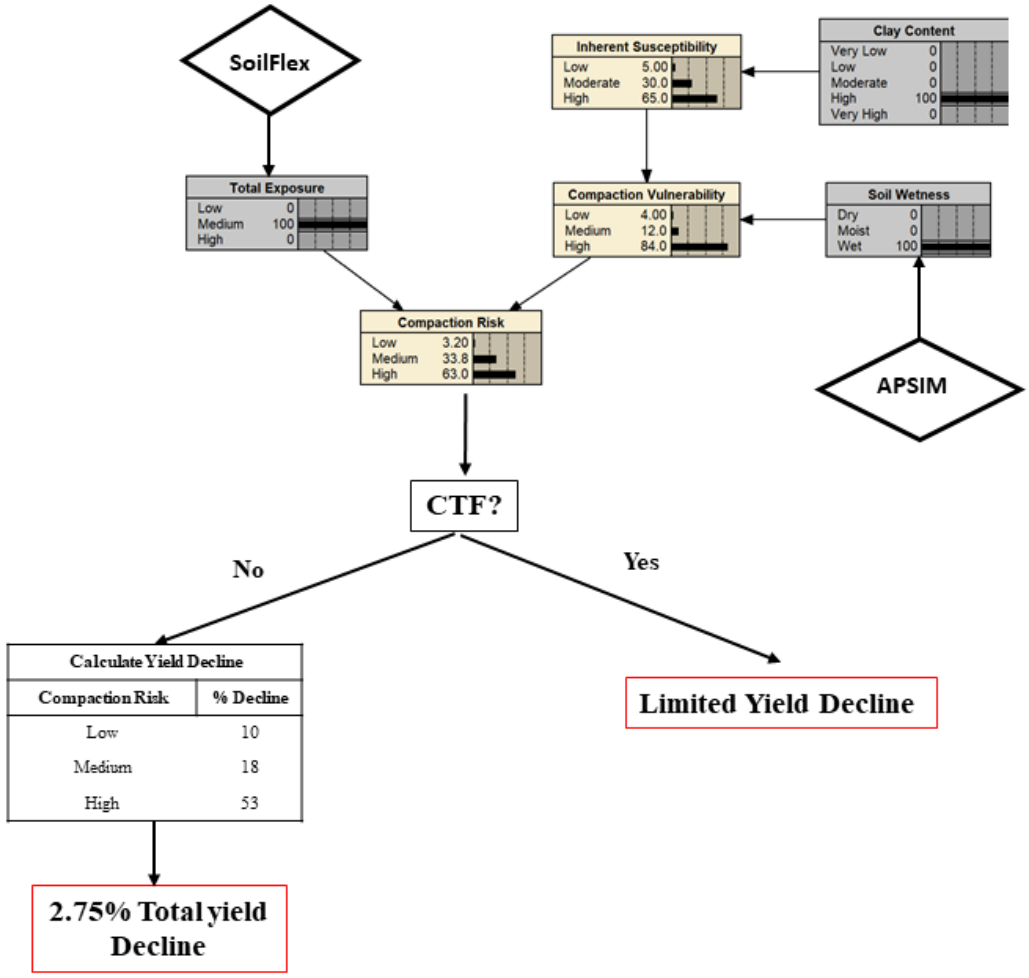


Figure 7.2 Main structure of developed compaction induced yield decline model using an example of 'Medium' Total exposure, 'High' clay content and 'Wet' Soil wetness. In an RTF system, this resulted in an estimated 2.75% reduction in yield for a single traffic event. SoilFlex and APSIM models can be employed to parameterize Total Exposure and Soil Wetness

7.2.1.A. Inherent susceptibility CPT

The inherent susceptibility describes the influence of soil clay content to the susceptibility of compaction using a conditional probability table (CPT). This was modified from Troldborg et al. (2013) to only include soil texture information as a means of simplifying model parameterisation for end users, which were predominantly intended to be landholders. The clay content of a soil is the primary determinant of susceptibility to compaction on the basis of cohesion and adhesion being the mechanical forces controlling particle realignment (Kirby, 1991). As the clay content effectively equates to *inherent susceptibility* in our model, it was ensured that the *inherent susceptibility* CPT was trained using published data (Figure 7.3) (Larson et al., 1980; Smith et al., 2007a) describing the compressibility index (C_c) as a function of clay content. While this data pertains to a mixture of Australian and American data, it is important to note that it remains applicable as only the ubiquitous mechanical aspects of particle cohesion and adhesion are being considered.

The compressibility index of a soil is a good proxy to infer the *inherent susceptibility* as it describes how a soil compresses under loading conditions independent of moisture conditions (Gupta and Larson, 1982). Soil compressibility was categorised into 'Low', 'Medium' and 'High' using boundaries adopted from Smith et al. (1997), as presented in Table 7.2 The raw data was used to directly train the *Inherent susceptibility* node as this allowed for the variability within the data to be represented as uncertainty in the predictions. This is a key benefit of BBNs over mode empirical approaches. Moderate and high probabilities for moderate and high clay content categories were adjusted with the addition of expert opinion to better represent the relationship between clay content and *Inherent susceptibility*. The resulting CPT for *Inherent susceptibility* is presented in Table 7.3.

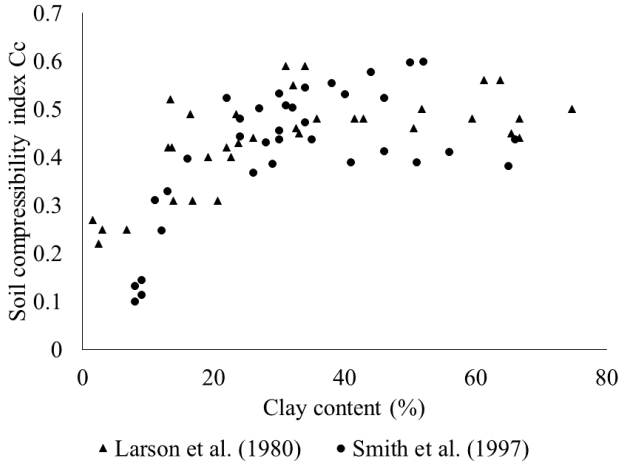


Figure 7.3 Published data describing clay content and compressibility index

Table 7.2 Boundaries of inherent susceptibility categories based on compressibility index.

Inherent susceptibility class	*Compressibility index range*
Low	<3
Moderate	3–4
High	>4

Table 7.3 Inherent susceptibility CPT as trained using published data.

Clay content	*Low*	*Moderate*	*High*
Very Low (0–20%)	55.6	22.2	22.2
Low (20–40%)	6.67	26.7	66.7
Moderate (40–60%)	5	15	80
High (60–80%)	5	15	80
Very High (80–100%)	9.09	18.2	72.7

7.2.1.B. Compaction vulnerability and compaction risk CPTs

The CPT for the *Compaction vulnerability* and *Compaction risk* node were adopted directly from Troldborg et al. (2013). The 'soil depth' node being removed as here we develop specific BNNs at each depth. The CPTs were developed using a published data, pedotransfer functions (PTFs), and expert opinion as described in Troldborg et al. (2013) (Table 7.4 Table 7.5).

Table 7.4 Compaction vulnerability CPT describing the probabilistic relationship between soil wetness, inherent susceptibility and compaction vulnerability.

Inherent susceptibility	*Soil wetness*	*Compaction vulnerability*		
		Low	*Medium*	*High*
Low	Dry	100	0	0
Low	Moist	75	25	0
Low	Wet	20	60	20
Moderate	Dry	75	25	0
Moderate	Moist	10	80	10
Moderate	Wet	10	60	30
High	Dry	50	50	0
High	Moist	0	50	50
High	Wet	0	0	100

Table 7.5 Compaction risk CPT describing the probabilistic relationship between total exposure, inherent susceptibility and compaction vulnerability.

Total exposure	*Compaction vulnerability*	*Compaction risk*		
		Low	*Medium*	*High*
Low	Low	100	0	0
Low	Medium	90	10	0
Low	High	0	75	25
Medium	Low	80	20	0
Medium	High	0	100	0
Medium	High	0	25	75
High	Low	0	100	0
High	Medium	0	25	75
High	High	0	0	100

7.2.2 Estimating yield decline

Yield decline due to soil compaction was estimated using published data (Table 7.6), which reported a yield decline due to a relative change in BD. The category boundaries for yield decline due to compaction risk were estimated using expert opinion, based on the correlations present within the collected data from the literature (Table 7.6). Yield reduction data were plotted against change in BD with a general relationship for increase in density to result in greater yield reduction (Figure 7.4). Using a linear model fitted to the data in conjunction with expert opinion, estimated magnitudes of yield decline associated with increases in BD were estimated (Table 7.7). The total yield impact, as a percentage of the total area trafficked, was subsequently calculated for a 9 m, 12 m and 18 m farming system. The assumption was made that the overriding factor in yield decline for crops, given increased density, was due to reduced infiltration via reduction in pore diameter. However, it is important to note that climatic, management, and biophysical constraints may also have existed.

Table 7.6. Yield decline as a function of the relative change in soil BD taken from literature.

Author	*Relative BD increase (%)*	*Yield decrease (%)*	*Crop*
(Ishaq et al., 2003)	14.9	18	wheat
(Botta et al., 2007)	10.03	19	soybean
	12.8	26.6	Soybean
	12.4	28.7	soybean
(Chan et al., 2006)	16	66	canola
(Sedaghatpour et al., 1995)	7	10.8	wheat
	9.8	12.3	faba bean

	18	3.5	wheat
	54	19.11	medic
(Chamen and Longstaff, 1995)	18	25	wheat
(Abu-Hamdeh, 2003)	3.5	17.6	corn
	5	25	corn
(Lipiec et al., 2003)	17.4	43	wheat
	8	4.5	barley
	21.8	21.2	barley
(Kulkarni et al., 2010)	8	55	cotton
	10	9	cotton
	14.5	37	cotton
	15.5	78	cotton
	16.2	29	cotton
	17.9	82	cotton

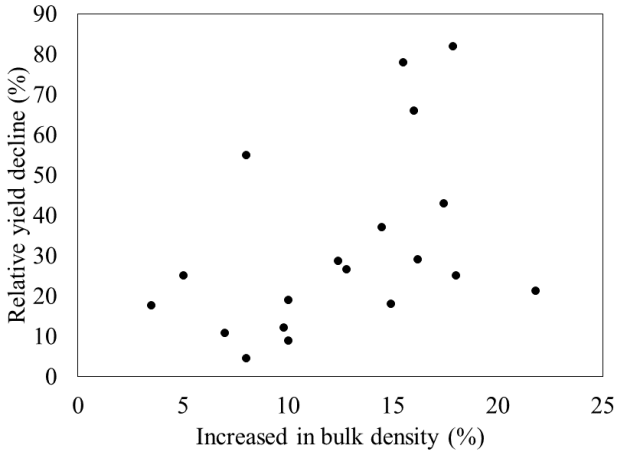


Figure 7.4 Published data describing the correlation between an increased in BD and relative yield decline.

Table 7.7 Estimated yield decline due to compaction risk. Paddock total yield declines calculated for 9, 12 and 18 m farming system implement widths. Tyre width taken to be 0.5 m for all traffic vehicles.

Compaction Risk	*BD increase (%)*	*Yield decrease*	*Paddock total decrease (%) (9 m system)*	*Paddock total decrease (%) (12 m system)*	*Paddock total decrease (%) (18 m system)*
Low	<5	10	1.11	0.833	0.556
Medium	5–15	18	2.04	1.53	1.02
High	>15	53	5.91	4.43	2.95

7.2.3 Simulations

The developed model was applied to a number of ApSoil profiles — the soil parameterization module internal to APSIM (Keating et al., 2003b) — to observe the temporal dynamics of compaction risk and yield decline, due to RTF, within key broad-acre cropping regions in Northern NSW, Australia. Whilst the compaction risk model is capable of estimating risk at any given depth, linking this to yield decline was only possible for profile estimates.

Therefore, yield declines were estimated using the mean total exposure, texture and soil moisture for the 0–60 cm profile. Estimates of compaction risk and yield decline were made within the months of May and November, to simulate standard planting and harvesting widows within the regions (Keating et al., 2003b; Sacks et al., 2010). This was achieved by modelling moisture dynamics using all available weather data for each site using APSIM (i.e. 01/01/1889 – 31/12/2017), and applying the developed model to estimate compaction risk on any given day. Averages were subsequently obtained for the months of May and November.

7.2.3.A. Machinery information

Stress profiles and associated compaction risk were estimated from 3 standard agricultural vehicles that were representative of the scale within the region of investigation, as examples. Details pertaining to their weight characteristics are provided in table Table 7.8.

Table 7.8 Machinery loading characteristics used in the calculation of yield decline. Data obtained from specification sheets provided by Deere and Company (2018a, 2018b, 2018c) Deere and Company (2018b)(Deere and Company, 2018c). Equal wheel load assumed.

Type	*Model*	*Tyre type*	*Tyre diameter (m)*	*Inflation pressure (kPa)*	*Total loaded weight (kg)*	*Load per wheel (kg)*
Sprayer	R4038	IF380/90R46	1.86	380	19,185	4,796
Tractor	8320R	480/80R50	2.05	241	17,195	4,299
Harvester	S670	600/70R28	1.57	240	31,561	7,890

7.2.3.B. Site information

Twenty ApSoil profiles representing a range of textures classes were selected to simulate the developed model. These were sourced directly from the ApSoil database within the APSIM initiative, and are presented in Table 7.9.

7.2.3.C. Soil moisture estimation

Daily soil moisture estimates were made using APSIM for the given sites. A wheat-on-wheat crop rotation was simulated to provide realistic differences in moisture between planting and harvest windows. Planting windows were provided as being between 15-May and 10-Jul, with harvest being set to automate once optimal. A depth of 50 mm of stored soil moisture was the minimum required to initiate sowing for any given cropping season. Weather data was

obtained from Silo Long-paddock (Queensland Government, 2018) and interpolated to the location of each site.

Table 7.9 Details of ApSoil profiles used in analysis. CWSP represents the Central West Slopes and Plains region of norther NSW, and NWSP represents the North West Slopes and Plains region of northern NSW. Texture class according to McDonald et al. (1998) and classified based on the dominant texture between 0–60 cm profile depth

ApSoil Site ID	*Cropping region*	*Latitude*	*Longitude*	*Dominant texture class*
1160	CWSP	-31.1006	148.2707	Light clay
702	CWSP	-33.5851	146.9305	Sandy loam
698	CWSP	-34.5029	147.1833	Sandy loam
693	CWSP	-34.029	147.3892	Sandy clay loam
247	CWSP	-31.0149	148.585	Sandy clay
199	CWSP	-32.495	147.174	Sandy clay loam
197	CWSP	-32.676	147.974	Medium clay
196	CWSP	-33.566	148.247	Light clay
1161	CWSP	-32.058	147.7091	Light clay
1014	NWSP	-30.3286	148.8642	clay
1016	NWSP	-29.6678	147.9993	clay
1279	NWSP	-29.1196	148.8318	Heavy clay
1285	NWSP	-29.2097	150.3275	Sandy clay
1288	NWSP	-28.7012	150.094	Medium clay
1290	NWSP	-31.3313	149.3857	Sandy clay
1292	NWSP	-30.9326	147.8779	Clay loam
1296	NWSP	-29.5832	150.3369	Heavy clay
1302	NWSP	-30.0211	148.1168	Medium clay
1308	NWSP	-31.3841	150.2258	Heavy clay

7.2.3.D. Farming system scenarios

A number of scenarios were tested to provide the likelihood of yield decreases, due to traffic of the soil under specific conditions. Each scenario represented a different farming system, depending on the level of matching of agricultural implement wheel-track widths and whether GPS guidance was utilised RTK or not. These scenarios were not established to give exact yield decline estimates, but instead to provide an indication on the yield effects of varied farming systems based on observations of relative yield decrease within the literature. The scenarios are as follows.

Scenario 1 – Conventional random traffic farming

This scenario represents conventional RTF where the direction of working is in a circular fashion around the field. In this situation, it is assumed each traffic event has essentially random placement of wheel tracks. Consequently, it is a reasonable assumption that with a 12

m implement and no reliable use of GPS guidance, 100% of field area will be trafficked within 5 years of cropping (Lipiec et al., 2003; Tullberg et al., 2007).

Scenario 2 – Unidirectional random traffic farming with two mismatched wheel tracks

This scenario represents RTF where all equipment passes are unidirectional, from one end of the field to the other without any cross working of the field. Implement widths are 36 m sprayer (3 m centres), 12 m planter (2 m centres) and 9 m harvester (3 m centres), resulting in two sets of mismatched wheel tracks with varied implement widths. Machine guidance is not RTK, therefore it is assumed annual wheel track drift is up to ±0.5 m from the centre of the initial pass between traffic events. Consequently, after 5 years it is calculated that 77.8% of field area is trafficked.

Scenario 3 – Controlled traffic farming (CTF) and no RTK

This scenario represents CTF where all equipment passes are unidirectional and implement widths remain as multiples of 12 m (i.e. 36 m sprayer, 12 m planter and 12 m harvester). However, machine guidance is not RTK. Therefore, it is assumed annual wheel track drift is up to ±0.5 m between traffic events. Consequently, after 5 years it is calculated that 16.7% of field area is trafficked.

Scenario 4 – Unidirectional random traffic farming with mismatched wheel tracks and RTK

This scenario represents a situation where all equipment passes are unidirectional with implement widths at multiples of 12, however with mismatched wheeltracks (i.e. 36 m sprayer on 4 m centres, 12 m planter on 2 m centres and 12 m harvester on 3 m centres). Consequently, after 5 years it is calculated that 19.4% of field area is trafficked.

Scenario 5 – True controlled traffic farming

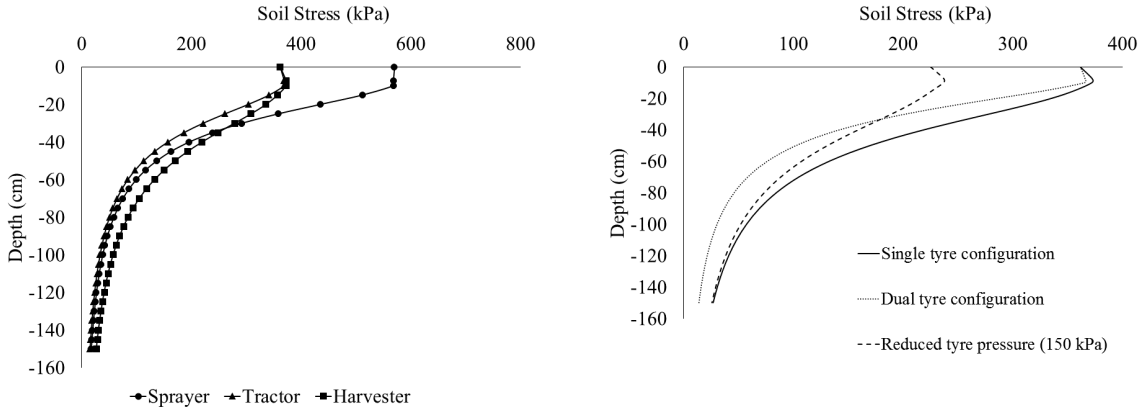
This scenario represents a true-CTF where all implement widths are sets of 12 m (i.e. 36 m sprayer, 12 m planter, 12 m planter). No wheel track drift is assumed due to RTK guidance (<2.0 cm accuracy). Consequently, at any point in time it is calculated that 8.3% of field area is trafficked.

7.3 Results

7.3.1 Vehicle stress profiles

Stress profiles were calculated for each agricultural vehicle and are presented in Figure 7.5. For each vehicle, soil stress within the top 35 cm was substantial, with all estimates being >120

kPa at 35 cm and >350 kPa at the surface, approaching 600 kPa for the sprayer. Therefore, all three vehicles exhibited high *total exposure*. Whilst the spraying vehicle occupied the largest surface loading conditions and soil stress within the 0–20 cm topsoil layer, the relative magnitudes of the other vehicles within this depth remained extreme. This suggests that whilst a reduction in surface loading conditions are achieved for some vehicles, their relative magnitudes do not warrant a change in compaction risk for topsoil conditions. Below 20 cm, stress induced by the harvesting vehicle remained the largest, within some categorical differs in exposure being detected within the 40–80 cm soil depth (Figure 7.5). Furthermore, when assessing the categorical stress distributions of each vehicle (Figure 7.6), the median *total exposure* state within the 0–60 cm range remained as 'high' for all vehicles. Reducing tyre inflation pressure or moving to a dual wheel configuration for the harvester did reduced the depth of 'high' *total exposure (*Figure 7.5*B))*, however at the cost of increased trafficked area.



(a) (b)

Figure 7.5 Distribution of soil stress below the surface for 3 agricultural vehicles (a) and beneath harvester loading (b) with dual tyre configuration and single tyre configuration with pressured reduced to 150 kPa

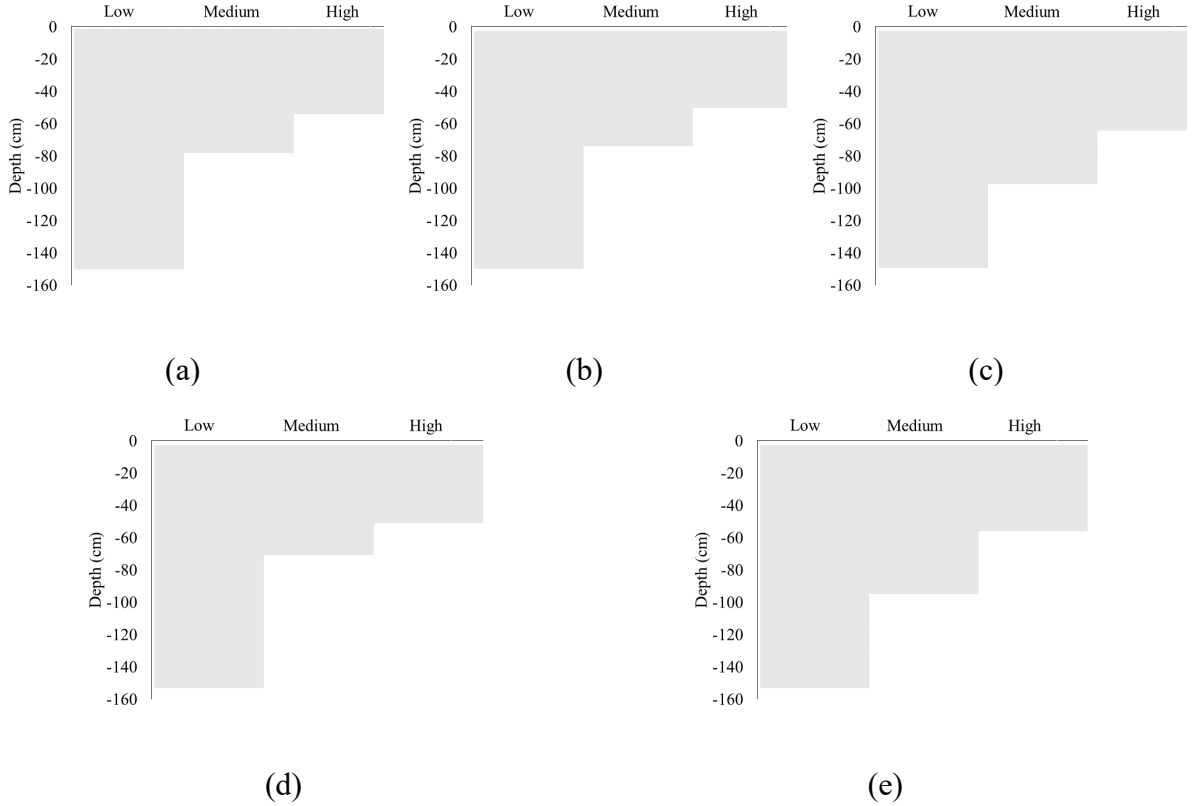


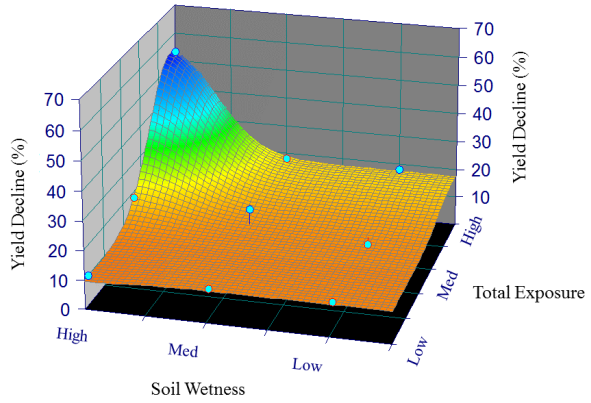
Figure 7.6 Total exposure profile estimates for a sprayer (a), tractor (b), harvester (c), harvester with dual tyres (d) and harvester with reduced tyre inflation pressure (150 kPa) (e).

7.3.2 Relationship of yield decline with soil wetness and total exposure

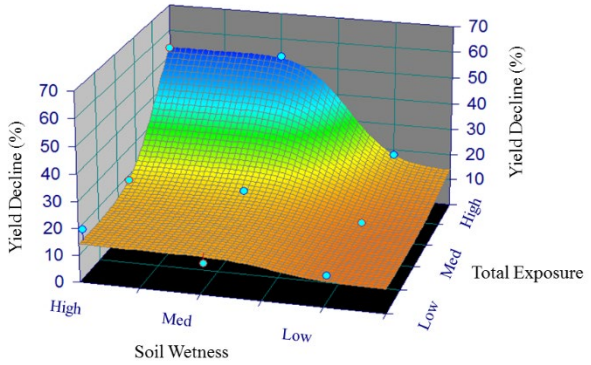
The response of yield decline for changes in soil wetness and total exposure for 3 different textures classes are presented in Figure 7.7. In general, increasing exposure, *soil wetness* and *clay content* result in an increased yield decline. For the very low *clay content* texture class (i.e. <15%), the magnitude of yield decline is only severe at high *total exposure* and *soil wetness* states, with lower *yield decline* state achieved toward the medium and low states of each node. However, as *clay content* is increased, the estimated *yield decline* for equivalent *total exposure* and *soil wetness* states increases significantly. T*otal exposure* has a greater influence on *yield decline* over *soil wetness*, with larger decline achieved at 'high' total exposure conditions over the equivalent decline for 'high' soil wetness.

The lack of a depth variable linked to yield confounds yield decline somewhat, but is consistent with a situation where soil moisture is sufficient for high compaction within the top 60 cm, which should be expected to result yield decline to some degree (expert opinion and literature value based CPT). The fact *total exposure* is the dominant node determining compaction risk highlights a serious industry concern – analytically demonstrated in Figure 7.6

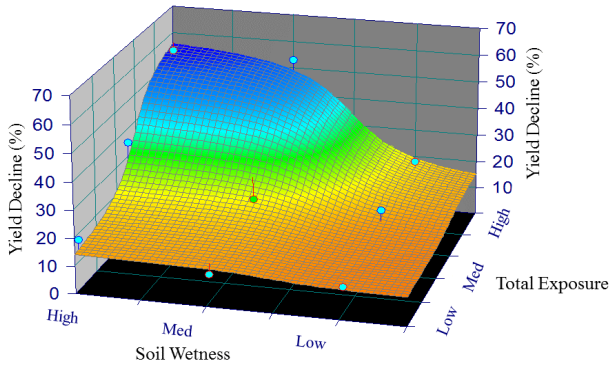
– whereby the majority of modern in-field agricultural vehicles result in ‘high’ total exposure due to their wheel-load and applied stress.



Very low clay content



Moderate clay content



Very high clay content

Figure 7.7 Fitted yield decline surface plots to categorical adjustments of soil wetness and total exposure nodes on the BBN. Yield decline is for the total decline directly underneath the wheel. Clay content corresponds to that presented in Table 7.3.

The difference in yield decline between 'moderate' and 'very high' clay contents is marginal, with the greatest difference being observed for 'high' *soil wetness* and 'medium' *total exposure*. This is a function of compressibility index, whereby *inherent susceptibility* increases rapidly from the low state towards the moderate state, then asymptotes at the high state (Table 7.2). That is, *compaction risk* and associated *yield decline* are initially highly sensitive to *clay content* increase, with *soil wetness* and *total exposure* conditions being more important at high *clay content*.

7.3.3 Single-pass yield declines

Using soil moisture calculations obtained from APSIM for a wheat-on-wheat cropping scenario, yield declines were estimated for 6 ApSoil sites exhibiting different textural classes (Table 7.1). For a given traffic event, potential yield declines were estimated based on the soil moisture conditions for individual days using the 128 years of simulated data. Key yield decline statistics for the months of May and November are presented in Table 7.10 for the 6 sites. As an example, for site 702, when soil wetness is classed as 'medium' and traffic is undertaken, an estimated yield decline of 2.44% of the paddock is likely, for a single traffic pass. These conditions however only present 11.8% of the time in May, and 5.7% of the time in November. For the site, 'low' soil wetness conditions are more likely, and represent 85% and 94% of the observations for May and November.

In general, as clay content increases across the sites, the likelihood of larger yield declines increases also, with 'high' soil moisture conditions being presented 91% and 90.3% for the 'very high clay content' sites (Table 7.10) Across all sites, reduced moisture conditions and subsequently reduced yield decline become more likely within the harvest month of November, in comparison to May. These dryer conditions are the result of water uptake from the planted crop which dries the profile during the growing season.

To understand the relative difference in compaction risk between regions, the estimated yield decline for a single traffic event on virgin soil was calculated on regional basis for planting traffic in May (Figure 7.8) and harvest traffic in November (Figure 7.9). This was achieved using the same simulated dataset described above. At each location, the likelihood of occurrence is based on the worst possible compaction state and yield decline. As an example, for the most Northern site, Moree, the most serve yield decline due to May planting traffic is estimated at 3.4% (based on a 12 m implement) for a single pass. This is observed 80-100 % of the time, based on soil moisture conditions.

Table 7.10 Yield decline severity and likelihood of occurrence based on site characteristics and simulated soil moisture conditions for selected ApSoil sites representing a range of clay content states.

ApSoil site	*Clay content*	*Soil wetness*	*Yield decline (%)*	*Occurrence (%)*	
				May planting	*November harvest*
		Low	1.76	85	94
702	Low	Med	2.44	11.8	5.7
		High	3	0	0
		Low	1.76	11.1	37.7
1292	Low	Med	2.45	70	51
		High	3.1	15.8	9.3
		Low	1.93	3	14.6
196	Medium	Med	2.96	20	33
		High	3.26	76	50
		Low	1.92	3.9	18.1
1161	Medium	Med	2.96	35.5	36.8
		High	3.25	57.4	42.5
		Low	1.98	2	9
1279	High	Med	3.13	4	30
		High	3.45	91	60
		Low	1.99	0	0
1016	High	Med	3.13	6.4	31.4
		High	3.45	90.3	65.4

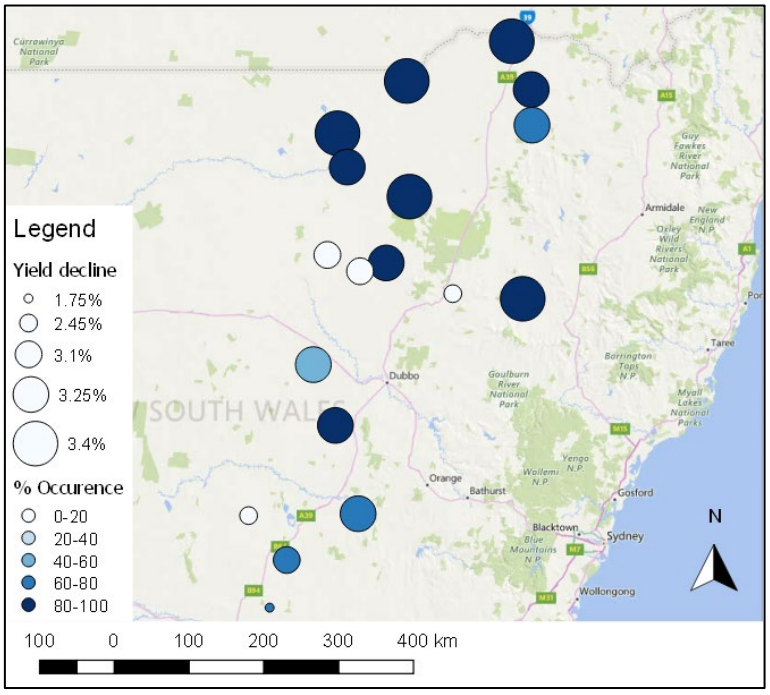


Figure 7.8 Yield decline risk map due to a single pass of planting traffic in May Yield decline is for the site total, based on a 12 m swathe width

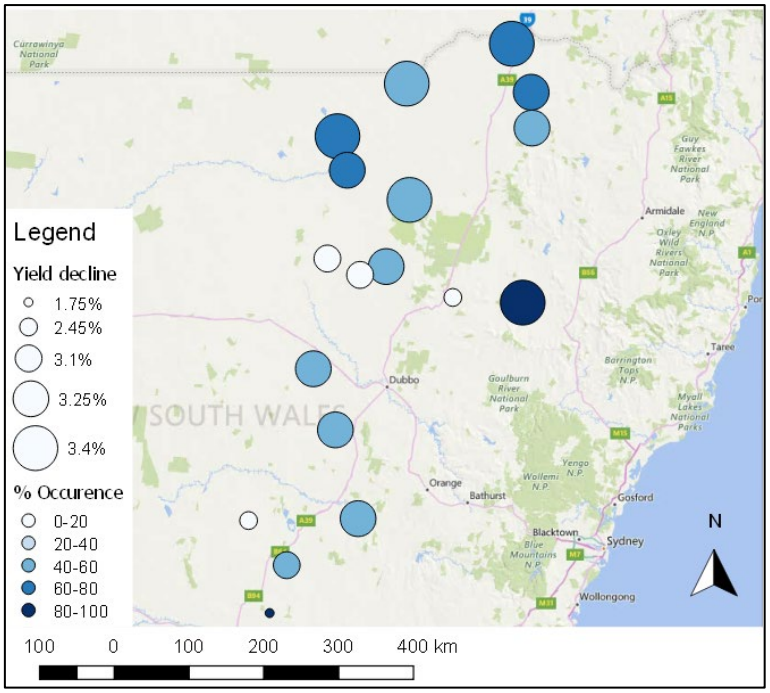


Figure 7.9 Yield decline risk map due to a single pass of harvest traffic in November. Yield decline is for the site total, based on a 12 m swathe width

7.3.4 Effects of RTF on yield decline

Agricultural vehicle/implement traffic systems with reduced consistency in wheel track width and/or failure to utilise RTK guidance result in an increased total area of soil receiving traffic. Consequently, a greater yield decline is probable based upon observed yield decline within the literature and the subsequent incorporation of this into the probabilistic model (Table 7.11). Yield reduction is greatest at high clay contents and high moisture conditions, which should be expected from the results presented above. The best case scenario occurs for true CTF farming (Scenario 5), with equal wheel track widths and single tyres; a yield decline in the order of 4.21% is likely for *very high clay* soils. This value should remain the benchmark, unless the frontage width of all implement can be increased to decrease the number of CTF tracks within the field. Comparatively, for the same soil, Scenario 4 with two mismatched wheel tracks using RTK guidance results in an estimated yield decline of 7.00%, highlighting the importance of implement matching. A CTF system without RTK guidance — annual drift in wheel tracks ±0.5 m from the centre of the initial pass — increases yield decline, beyond Scenario 4, up to 8.46% when implement widths are matched (Scenario 3). The effect of this drift is exacerbated in instances with mismatched wheel tracks (Scenario 2), with a yield decline of up to 19.7% being estimated for a high clay content soil, and increasing to a 50.8% yield penalty for conventional concentric agricultural vehicle pass RTF systems.

Table 7.11 Estimated yield decline due to modern day agricultural vehicles for each traffic scenario at varied clay content and moisture conditions.

Clay content	*Moisture*	*Scenario (area trafficked)*				
		1 (100%)	*2 (77.8%)*	*3 (16.7%)*	*4 (19.4%)*	*5 (8.3%)*
Very Low Clay	dry	22.4	17.4	3.73	4.34	1.86
	moist	33.9	26.4	5.65	6.57	2.81
	wet	44.3	34.4	7.38	8.58	3.67
Medium clay	dry	25.2	19.6	4.20	4.89	2.09
	moist	42.5	33.1	7.09	8.25	3.53
	wet	47.5	37.0	7.92	9.22	3.94
Very high clay	dry	28.7	22.4	4.79	5.58	2.39
	moist	45.3	35.3	7.56	8.79	3.76
	wet	50.8	39.5	8.46	9.85	4.21

7.3.5. Compaction risk profiles

Compaction risk profiles for three sites are presented in Figure 7.10. These were taken as the average conditions observed in May and November between1997–2017. These profiles descried depth-specific compaction risk during times of likely field operations (planting in June using a tractor, and harvesting in November using a harvester, and tractor-pulled grain haul out bin). For the two sites that have a low – moderate texture class through the profile (id 702 and 196), the compaction risk does not change between May and November under tractor loading conditions. For the high clay content site however (id 1279), compaction risk in the 30–120 cm soil layers is lower in November in comparison to May, suggesting the water extraction of the crop was sufficient to lower the compaction risk of tractor traffic. For harvest traffic however, the risk remained consistent in the 0–90 cm for November traffic in comparison to May tractor traffic.

When considering the relative differences between tractor and harvest traffic for November, higher risk conditions were generally observed deeper in the profile, in comparison to tractor traffic, likely due to the increased stress. Interesting, for the moderate clay content site (id 196), the risk of soil compaction was greatest in the 30–60 cm layer, suggesting that subsoil compaction was more likely than surface compaction. For the high clay content site (id 1279), soil compaction remains ‘medium’ for deep layers (> 50 cm), which may suggest the model is not well calibrated to wet soil moisture conditions at very high clay contents experiencing low exposure conditions.

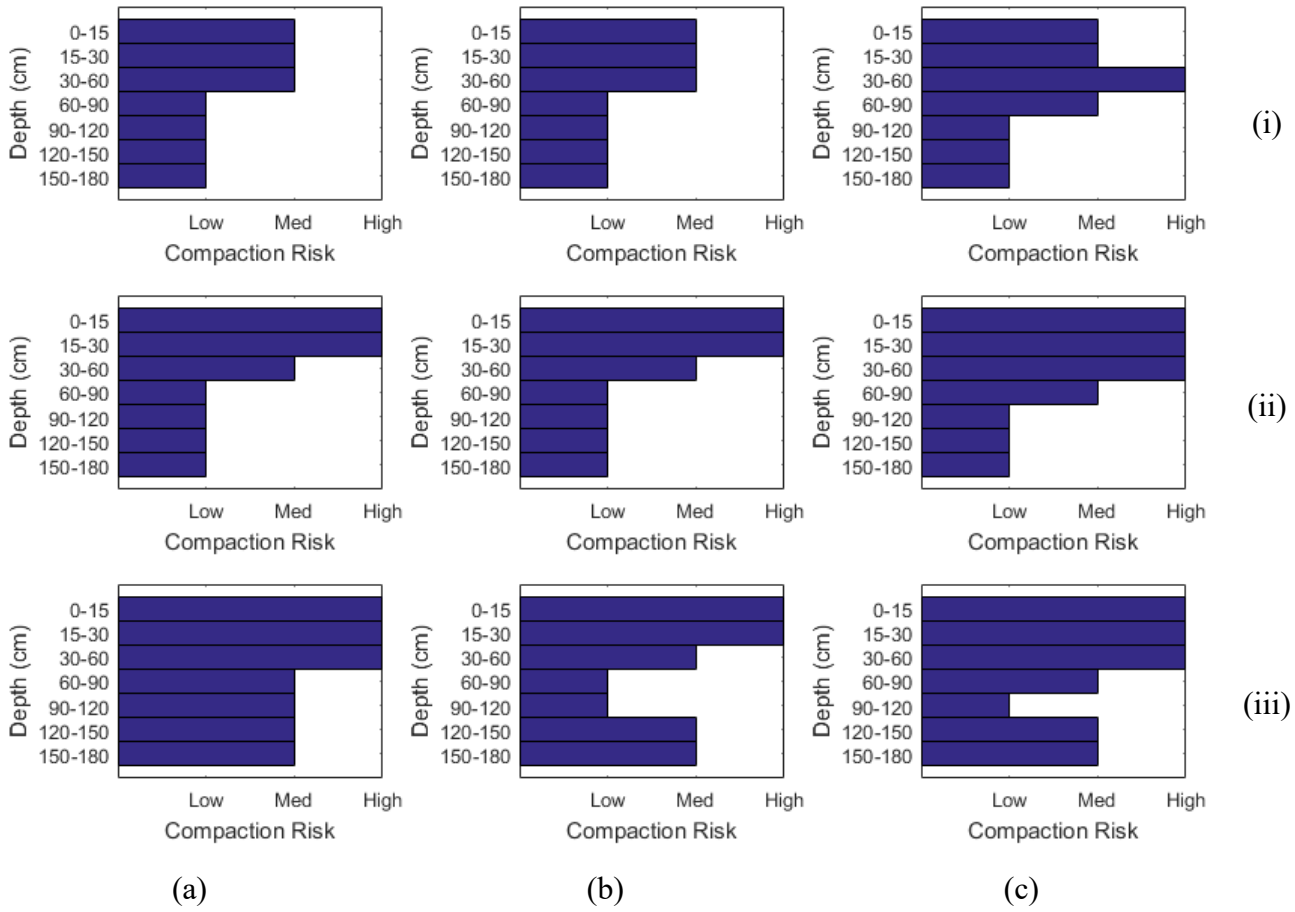


Figure 7.10 Total exposure profiles for three sites with varied texture classes (very low clay content (i), moderate clay content (ii) and very high clay content (iii)) based on the average of conditions observed in May under tractor loading (a) and November under tractor (b) and harvesting loading (c) conditions. ApSoil site ids for the sites are as follows: 702(i), 196 (ii) and 1279 (iii).

7.4. Discussion

7.4.1 Evaluation of BBN model

The intention of the BBN compaction model is to provide agricultural practitioners with the investigative capability for the magnitude of cost to benefit of traffic management systems, whereby soil compaction is a risk and yield penalty provided as probable relative reduction based on literature observations. Clear mandates on conversion to CTF within agricultural systems is either largely missing (Bennett et al., 2019; Tullberg et al., 2018) or the benefits are not clearly articulated to growers. The developed model is a useful tool in the assessment of soil compaction as a function of climate, field traffic conditions and inherent site characteristics in context of lost grain production. This enables the interactions between key compaction drivers to be investigated, therefore providing opportunity to realise the sensitivity of compaction effects to management practices and decisions. Bennett et al. (2015b) and Tullberg

et al. (2007) indicate that the adoption of CTF is less than one third of the Australian cropping sector, meaning that the capability to weight up management scenarios, such as RTF, varied implement widths or field operations at less than ideal moisture conditions is an important advancement. Both predictive and diagnostic reasoning can be achieved using the model. Predictive reasoning can be useful in identifying the likely impact of a given operation, therefore providing the basis for cost-benefit assessment of field operation timing. Alternatively, diagnostic reasoning can be performed with the tool, meaning the conditions required to achieve a defined level of yield decline or compaction risk can be identified. This allows the user the capability to systematically adjust the critical control points of the system to identify the necessary system-states required to achieve a desired outcome. As an example, for high clay content soils (i.e. >35%) under 'medium' total exposure conditions, the user would be required to wait for soil moisture conditions to reduce to 'low' (i.e. <20%). Alternatively, total exposure would be required to be reduced to 'low' (<60 kPa), to achieve the same outcome.

The developed BBN also allows for the assessment of compaction at various scales, dependent on the spatial resolution of information that's available. Here we have presented point-based estimates of the risk of compaction and subsequent yield decline for point-locations across a large geographic region (Figure 7.8 and Figure 7.9). The BBN has identified the key regions where compaction risk is frequently high, including the Spring Ridge site in the Liverpool Plains of NSW (ApSoil id 1308). This region is dominated by high clay content soils (Young et al., 2009) and a large average annual rainfall (634 mm), meaning soil moisture conditions are frequently conducive to high soil compaction risk. Alternatively, the model can be applied to map compaction-risk at the sub-field scale, if clay % and soil moisture estimates were available. For growers using this tool with access to this data, it would enable them to identify regions of high and low risk, therefore providing opportunity to spatially prioritise field operations of minimising detrimental soil compaction.

7.4.2 The value proposition for CTF

Despite the CTF philosophy being present for almost 4 decades (Taylor, 1983; Voorhees et al., 1979), it's uptake in industry remains limited, despite countless literature highlighting its benefits (Antille et al., 2016a; Bartimote et al., 2017; Bennett et al., 2019; Bennett et al., 2017; Kingwell and Fuchsbichler, 2011; McHugh et al., 2009; Roberton and Bennett, 2017; Tullberg et al., 2007). This is due to the perceived yield benefits in comparison to the cost of machinery upgrade (Bennett et al., 2017). The BBN compaction model developed provides the capability

to compare the yield effects of RTF farming with various levels of implement wheel track mismatching, based on a collation of data from literature. While the yield decline estimated is not expected to accurately represent the exact yield decline, it provides an evidence based relative estimate of the probable impacts of RTF. The approach presented is pragmatic to overcome the lack of uncompacted benchmark data. Furthermore, it provides the ability to test management scenarios without having to implement them, overcoming the requirement for site-specific, nuanced management data affecting compaction that is rarely collated and collected, but is expected to greatly affect the controlling factors of soil compaction (i.e. the inability to analytically model management and system complexity). Using true CTF as the benchmark (minimised 4.21% decline), in comparison to Scenario 2 — unidirectional RTF with two mismatched wheel tracks and inconsistent implement widths — and Scenario 4 — unidirectional random traffic farming with mismatched wheel tracks and RTK (matched implement widths) — it is observed that the CTF system offers significant advantage. It is a fair assumption that the *wet* soil moisture condition will be present at some point of traffic within a paddocks traffic history (Braunack and Johnston, 2014) Therefore, these conditions should be considered when exploring the economics surrounding total yield decline. For *very high clay* soils, these scenarios result in a potential yield decline (relative to CTF) of 35.3% and 5.64% respectively. For an arbitrary farm size of 1000 ha, an average wheat price of $241/t (ABARES, 2019) and an average wheat yield of 2 t/ha (Australian Bureau of Statistics, 2015), this reduction equates to $170,146 and $27,184 respectively, for a single cropping season. Considering the average total-farm CTF upgrade cost of $147,000 (Bowman, 2008), upgrade to CTF should be considered to be recovered within <5 years for scenario 4, and within a single cropping system for scenario 2.

7.4.3 Towards depth-based risk assessment

Whilst the developed model can estimate the compaction risk within specific depth-layers, the lack of literature pertaining to depth-specific compaction effects on yield limits our ability to attribute yield decline to specific compacted layers. Whilst soil stress, and subsequently *total exposure* decreases with depth during loading, soil moisture and clay content generally increase with depth, therefore increasing the *inherent susceptibility* to compaction. As a result, the compaction risk may be more substantial within layers below the surface. Adding to this risk is the increased cost of compaction amelioration to depth (Håkansson and Reeder, 1994), which requires significant time and energy. The financial significance to subsoil

compaction is therefore large, not only from a yield decline perspective, but also from an amelioration cost perspective. Future work should seek to quantify yield declines with depth-specific information in order to build a more robust compaction management decision tool.

7.4.4 The efficacy of adjusting machine parameters to reduce compaction risk

Based on SoilFlex input for tyre characteristics and resulting stress-state distribution, it was demonstrated that increasing the tyre contact area by reducing tyre inflation pressure, or increasing the number of tyres, for the axle loads of the modern agricultural vehicles (Figure 7.5 andFigure 7.6), did not affect the risk of compaction. Whilst some merit exists to reduce soil stress by making the adjustments to agricultural vehicles (Antille et al., 2016b; Keller and Arvidsson, 2004), the magnitude at which stress is decreased is not sufficient to reduce the risk compaction occurring at the specified depths. This is because the soil stress continues to exceed that of precompression stresses (Chamen et al., 2003; Horn and Fleige, 2003), despite an increased loading area, therefore resulting in soil failure and subsequently severe, and often permanent, structural changes (Alaoui et al., 2011; Lipiec et al., 2012). Adjusting tyre configurations of agricultural vehicles is unlikely to avoid yield declines due to traffic, and the model provides the means to demonstrate this to practitioners.

Yield decline can be significantly reduced by achieving 'Low' *total exposure* conditions (e.g. 1.36% and 1.05% for 'high' and 'very low' clay contents at 'medium' soil wetness in a 12 m system). However, for the machinery investigated here, 'low' *total exposure* conditions are never achieved within the major rooting depth (0–50 cm), despite wheel loads of the tractor being 46% less than that of the harvester. Subsequently, *total exposure* conditions remained 'high' for all machines in the calculation of yield decline (i.e. average stress in the 0–60 cm profile layer). In fact, wheel loading is required to be reduced below 1300 kg to reduce *total exposure* to a 'medium' category. Håkansson (1990) makes the assertion that vehicle wheel loads should never exceed a surface stress of 200 kPa, stating that such vehicles that do should never be allowed onto agricultural fields. However, given the scale of current field operations, equipment below this magnitude is rarely used, which is thus true for agricultural vehicles with wheel load <1300 kg. Furthermore, attempting to achieve small reductions in weight of common agricultural vehicles is unlikely to significantly improve yield decline, given the already substantial size of the machines. Therefore, without significantly reducing the wheel load mass, which requires smaller, lighter machines machines, or an increase in the number of

axles, the *total exposure* can effectively only be lessened by confining the spatial traffic impact via CTF management, which is consistent with the assertions within the literature (e.g. Antille et al., 2016a).

7.4.5 Opportunities for qualitative assessment of soil constraints

Presented here is an approach that integrates quantitative data sources with qualitative information to provide inference toward soil health and its effects of yield. The incorporation of qualitative management information is a key benefit of this approach, as this information can be highly influential on system function (Aalders, 2008; Sattler et al., 2010). Furthermore, management strategies can be largely unique to a specific site, based on the land owner's previous experiences and current objectives for management (Bennett and Cattle, 2013; Bennett and Cattle, 2014). These strategies may subsequently effect the way in which constraints are presented; e.g. the effects of surface sodicity are increased without stubble retention (Alzubaidi and Webster, 1982; Yaduvanshi and Sharma, 2008). Therefore, capturing this information becomes prudent in the assessment of soil constraints and their amelioration.

7.5 Conclusion

The work presented here has developed a novel tool to assess soil compaction risk semi-quantitatively using a BBN approach. This has allowed for the effects of surface loading, soil clay content and soil moisture on compaction risk to be explored. The tool was able to broadly estimate the yield impacts due to various agricultural traffic scenarios. Subsequently, the financial consequences of failing to adopt CTF management within an agricultural enterprise was identified, suggesting that the cost of CTF conversion could be recovered in a short period of time (<5 years). Tools such as these are required in industry to broadly describe the impacts of RTF farming to aid in the uptake of CTF.

Importantly, the work presented here has showcased the use of an alternative approach to investigate constraint-yield interactions that are difficult to assess quantitatively using empirical models. The developed BBN approach is capable of providing inference be merging quantitative and qualitative management information that is highly influential on the impacts of compaction. Consequently, this approach provides opportunity in data limiting environments and should be considered a useful tool where empirical models struggle.

7.6 References

Aalders, I., 2008. Modeling land-use decision behavior with Bayesian belief networks. Ecology and Society 13(1).

ABARES, 2019. Weekly Australian Climate, Water and Agricultural Update. Department of Agriculture and Water Resouces.

Abu-Hamdeh, N.H., 2003. Compaction and subsoiling effects on corn growth and soil bulk density. Soil Science Society of America Journal 67(4), 1213-1219.

Alaoui, A., Lipiec, J., Gerke, H., 2011. A review of the changes in the soil pore system due to soil deformation: A hydrodynamic perspective. Soil and Tillage Research 115, 1-15.

Alzubaidi, A., Webster, G., 1982. Effect of tillage in combination with chemical amendments on reclamation of a solonetzic soil. Canadian Journal of Soil Science 62(4), 641-649.

Antille, D.L., Bennett, J.M., Jensen, T.A., 2016a. Soil compaction and controlled traffic considerations in Australian cotton-farming systems. Crop and Pasture Science 67(1), 1-28.

Antille, D.L., Bennett, J.M., Jensen, T.A., Roberton, S.D., 2016b. The influence of tyre inflation pressure on soil compaction caused by the John Deere 7760. An impact assessment framework for harvesting technologies in cotton: Management considerations forthe John Deere 7760 National Centre for Engineering in Agriculture Publication 1004960/16/1, USQ, Toowoomba.

Australian Bureau of Statistics, 2015. Agricultural Commodities, Australia. cat. no. 7121.0.

Bailey, A., Raper, R., Johnson, C., Burt, E., 1995. An integrated approach to soil compaction prediction. Journal of Agricultural Engineering Research 61(2), 73-80.

Bartimote, T., Quigley, R., Bennett, J.M., Hall, J., Brodrick, R., Tan, D.K., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: II. Economic and physiological analysis. Soil and Tillage Research 168, 133-142.

Bennett, J.M., Cattle, S., 2013. Adoption of soil health improvement strategies by Australian farmers: I. Attitudes, management and extension implications. The Journal of Agricultural Education and Extension 19(4), 407-426.

Bennett, J.M., Cattle, S., 2014. Adoption of soil health improvement strategies by Australian farmers: II. Impediments and incentives. The Journal of Agricultural Education and Extension 20(1), 107-131.

Bennett, J.M., Cattle, S., Singh, B., 2015. The efficacy of lime, gypsum and their combination to ameliorate sodicity in irrigated cropping soils in the Lachlan Valley of New South Wales. Arid Land Research and Management 29(1), 17-40.

Bennett, J.M., Roberton, S., Marchuk, S., Woodhouse, N., Antille, D., Jensen, T., Keller, T., 2019. The soil structural cost of traffic from heavy machinery in Vertisols. Soil and Tillage Research 185, 85-93.

Bennett, J.M., Roberton, S.D., Jensen, T.A., Antille, D.L., Hall, J., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: I. Heavy machinery impact on the soil resource. Soil and Tillage Research 168, 143-154.

Botta, G.F., Pozzolo, O., Bomben, M., Rosatto, H., Rivero, D., Ressia, M., Tourn, M., Soza, E., Vazquez, J., 2007. Traffic alternatives for harvesting soybean (Glycine max L.): effect on yields and soil under a direct sowing system. Soil and Tillage Research 96(1-2), 145-154.

Bowman, K., 2008. Economic and environmental analysis of converting to controlled traffic farming, Proceedings 6th Australian Controlled Traffic Farming Conference, pp. 12-14.

Braunack, M., Johnston, D., 2014. Changes in soil cone resistance due to cotton picker traffic during harvest on Australian cotton soils. Soil and Tillage Research 140, 29-39.

Chamen, T., Alakukku, L., Pires, S., Sommer, C., Spoor, G., Tijink, F., Weisskopf, P., 2003. Prevention strategies for field traffic-induced subsoil compaction: a review: Part 2. Equipment and field practices. Soil and Tillage Research 73(1), 161-174.

Chamen, W., Longstaff, D., 1995. Traffic and tillage effects on soil conditions and crop growth on a swelling clay soil. Soil Use and Management 11(4), 168-176.

Chan, K., Oates, A., Swan, A., Hayes, R., Dear, B., Peoples, M., 2006. Agronomic consequences of tractor wheel compaction on a clay soil. Soil and Tillage Research 89(1), 13-21.

Dang, Y., Dalal, R., Mayer, D., McDonald, M., Routley, R., Schwenke, G., Buck, S., Daniells, I., Singh, D., Manning, W., 2008. High subsoil chloride concentrations reduce soil water extraction and crop yield on Vertosols in north-eastern Australia. Australian Journal of Agricultural Research 59(4), 321-330.

Deere and Company, 2018a. Product brochure: 4 Series Self-Propelled Sprayers.

Deere and Company, 2018b. Product brochure: 8R/8RT Series.

Deere and Company, 2018c. Product brochure: Combines.

Gupta, S., Larson, W., 1982. Modeling Soil Mechanical behavior During Tillage 1. Predicting tillage effects on soil physical properties and processes (predictingtilla), 151-178.

Håkansson, I., 1990. A method for characterizing the state of compactness of the plough layer. Soil and tillage research 16(1-2), 105-120.

Håkansson, I., Reeder, R.C., 1994. Subsoil compaction by vehicles with high axle load—extent, persistence and crop response. Soil and Tillage Research 29(2), 277-304.

Hamza, M., Anderson, W., 2005. Soil compaction in cropping systems: A review of the nature, causes and possible solutions. Soil and tillage research 82(2), 121-145.

Henriksen, H.J., Barlebo, H.C., 2008. Reflections on the use of Bayesian belief networks for adaptive management. Journal of Environmental Management 88(4), 1025-1036.

Horn, R., Fleige, H., 2003. A method for assessing the impact of load on mechanical stability and on physical properties of soils. Soil and Tillage Research 73(1-2), 89-99.

Ishaq, M., Ibrahim, M., Lal, R., 2003. Persistence of subsoil compaction effects on soil properties and growth of wheat and cotton in Pakistan. Experimental Agriculture 39(4), 341-348.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N., Meinke, H., Hochman, Z., 2003. An overview of APSIM, a model designed for farming systems simulation. European journal of agronomy 18(3), 267-288.

Keller, T., Arvidsson, J., 2004. Technical solutions to reduce the risk of subsoil compaction: effects of dual wheels, tandem wheels and tyre inflation pressure on stress propagation in soil. Soil and Tillage Research 79(2), 191-205.

Keller, T., Défossez, P., Weisskopf, P., Arvidsson, J., Richard, G., 2007. SoilFlex: A model for prediction of soil stresses and soil compaction due to agricultural field traffic including a synthesis of analytical approaches. Soil and Tillage Research 93(2), 391-411.

Kingwell, R., Fuchsbichler, A., 2011. The whole-farm benefits of controlled traffic farming: An Australian appraisal. Agricultural Systems 104(7), 513-521.

Kirby, J., 1991. Critical-state soil mechanics parameters and their variation for Vertisols in eastern Australia. Journal of Soil Science 42(3), 487-499.

Kulkarni, S., Bajwa, S., Huitink, G., 2010. Investigation of the effects of soil compaction in cotton. Transactions of the ASABE 53(3), 667-674.

Larson, W., Gupta, S., Useche, R., 1980. Compression of Agricultural Soils from Eight Soil Orders 1. Soil Science Society of America Journal 44(3), 450-457.

Lawes, R., Oliver, Y., Robertson, M., 2009. Capturing the in-field spatial–temporal dynamic of yield variation. Crop and Pasture Science 60(9), 834-843.

Lipiec, J., Hajnos, M., Świeboda, R., 2012. Estimating effects of compaction on pore size distribution of soil aggregates by mercury porosimeter. Geoderma 179, 20-27.

Lipiec, J., Medvedev, V., Birkas, M., Dumitru, E., Lyndina, T., Rousseva, S., Fulajtar, E., 2003. Effect of soil compaction on root growth and crop yield in Central and Eastern Europe. International agrophysics 17(2), 61-70.

Littleboy, M., Silburn, D., Freebairn, D., Woodruff, D., Hammer, G., 1989. PERFECT-A computer simulation model of Productivity Erosion Runoff Functions to Evaluate Conservation Techniques. Bulletin-Queensland Department of Primary Industries (Australia).

McDonald, R.C., Isbell, R., Speight, J.G., Walker, J., Hopkins, M., 1998. Australian soil and land survey: field handbook. CSIRO publishing.

McHugh, A., Tullberg, J., Freebairn, D., 2009. Controlled traffic farming restores soil structure. Soil and Tillage Research 104(1), 164-172.

Nuttall, J., Armstrong, R., 2010. Impact of subsoil physicochemical constraints on crops grown in the Wimmera and Mallee is reduced during dry seasonal conditions. Soil Research 48(2), 125-139.

Nuttall, J.G., Armstrong, R., Connor, D., 2003. Evaluating physicochemical constraints of Calcarosols on wheat yield in the Victorian southern Mallee. Australian Journal of Agricultural Research 54(5), 487-497.

O'sullivan, M., Henshall, J., Dickson, J., 1999. A simplified method for estimating soil compaction. Soil and Tillage Research 49(4), 325-335.

Orton, T.G., Mallawaarachchi, T., Pringle, M.J., Menzies, N.W., Dalal, R.C., Kopittke, P.M., Searle, R., Hochman, Z., Dang, Y.P., 2018. Quantifying the economic impact of soil constraints on Australian agriculture: A case-study of wheat. Land Degradation & Development 29(11), 3866-3875.

Queensland Government, 2018. SILO climate data.

Roberton, S.D., Bennett, J.M., 2017. Efficacy of delaying cotton defoliation to mitigate compaction risk at wet harvest. Crop and Pasture Science 68(5), 466-473.

Robertson, D., Wang, Q.J., 2004. Bayesian networks for decision analyses — an application to irrigation system selection. Australian Journal of Experimental Agriculture 44(2), 145-150.

Sacks, W.J., Deryng, D., Foley, J.A., Ramankutty, N., 2010. Crop planting dates: an analysis of global patterns. Global Ecology and Biogeography 19(5), 607-620.

Sattler, C., Nagel, U.J., Werner, A., Zander, P., 2010. Integrated assessment of agricultural production practices to enhance sustainable development in agricultural landscapes. Ecological Indicators 10(1), 49-61.

Sedaghatpour, S., Ellis, T., Hignett, C., Bellotti, B., 1995. Six years of controlled traffic cropping research on a red brown earth at Roseworthy in South Australia, Proceedings 1st National Controlled Traffic Conference, pp. 13-14.

Smith, C., Johnston, M., Lorentz, S., 1997. Assessing the compaction susceptibility of South African forestry soils. II. Soil properties affecting compactibility and compressibility. Soil and Tillage Research 43(3-4), 335-354.

Smith, C.S., Howes, A.L., Price, B., McAlpine, C.A., 2007. Using a Bayesian belief network to predict suitable habitat of an endangered mammal–The Julia Creek dunnart (Sminthopsis douglasi). Biological Conservation 139(3-4), 333-347.

Taylor, J.H., 1983. Benefits of permanent traffic lanes in a controlled traffic crop production system. Soil and Tillage Research 3(4), 385-395.

Troldborg, M., Aalders, I., Towers, W., Hallett, P.D., McKenzie, B.M., Bengough, A.G., Lilly, A., Ball, B.C., Hough, R.L., 2013. Application of Bayesian Belief Networks to quantify and map areas at risk to soil threats: Using soil compaction as an example. Soil and Tillage Research 132, 56-68.

Tullberg, J., Antille, D.L., Bluett, C., Eberhard, J., Scheer, C., 2018. Controlled traffic farming effects on soil emissions of nitrous oxide and methane. Soil and Tillage Research 176, 18-25.

Tullberg, J., Yule, D., McGarry, D., 2007. Controlled traffic farming—from research to adoption in Australia. Soil and Tillage Research 97(2), 272-281.

Voorhees, W., Young, R., Lyles, L., 1979. Wheel traffic considerations in erosion research. Transactions of the ASAE 22(4), 786-0790.

Yaduvanshi, N., Sharma, D., 2008. Tillage and residual organic manures/chemical amendment effects on soil organic matter and yield of wheat under sodic water irrigation. Soil and tillage research 98(1), 11-16.

Young, R., Wilson, B., Harden, S., Bernardi, A., 2009. Accumulation of soil carbon under zero tillage cropping and perennial vegetation on the Liverpool Plains, eastern Australia. Soil Research 47(3), 273-285.

8. General discussion, conclusions and future research directions

8.1 General discussion

Soil constraints are highly variable across the landscape (Dang and Moody, 2016; McBratney and Pringle, 1999) and exhibit complex interactions that impede the systematic diagnosis of their spatial effects on crop yield. Due to the perceived costs of data acquisition in Australian agriculture, the resolution of soil sampling is often insufficient to accurately identify and account for this variability when providing spatial advice for soil amelioration. Hence, the application of soil ameliorants are frequently highly inaccurate and inadequate, and often do not address the constraints presented, leading to over- and under-application of resources and failure to realise the yield potential. It is prudent to understand economics of increased data collection costs with consideration toward the benefit of improved agronomic advice. Balancing the cost-benefit of agronomic advice requires knowledge of spatial prediction errors of DSM approaches, and the interactions between site-specific soil function and crop yield. Investigations presented in this thesis are based on data and information from a single commercial agricultural site to examine in depth the use of DSM and pedometrics to optimise spatial constraint management. The true novelty of the work is to utilise the various methods to provide management recommendations on a spatial scale that allows variable rate management. Opportunities and limitations in the wider application of this framework are also discussed.

8.1.1. Optimised sampling density in relation to soil spatial variability

In chapter 4 of this thesis, the optimal sampling densities for constraint management were defined as 0.2 and 0.5 cores/ha for 0–20 cm and 0–60 cm treatment respectively. This was using an intensively samples site as 2.78 cores/ha. Whilst this sampling density is rarely acquired in dryland agriculture, this dataset for a single field provides the basis to explore how optimum densities can be identified at unvisited fields. The resource constrained sampling density is dependent upon the spatial variability of properties of interest, the budgetary constraint in practical implementation of the survey and the spatial prediction error acceptable to the user (Simbahan and Dobermann, 2006). Whilst methods exist to estimate this optimal density when some knowledge of the spatial variability of the target variable is already known (Bogaert and Russo, 1999; Lark,

2002; McBratney et al., 1981), this prior information is not readily available at the field-scale in agriculture. Whilst McBratney and Pringle (1999) provide an approach to overcome this using average variograms, the spatial nature of the target variable is often inherent to the specific site. Therefore, a metric is required to infer the variability of the target variable/s, prior to sampling, to estimate the likely optimal sampling density, considers the cost-benefit of data acquisition for soil amelioration purposes.

We propose the use of environmental covariates as a proxy estimate of the spatial variability within a given site (e.g. yield data, elevation data, proximally sensed data etc.), which has similarly been utilised for the construction of SSPFe models (Malone et al., 2018). However, in terms of the current work these environmental covariates may not be directly utilised in the resulting spatial prediction models. That is, while not used for spatial predictions (e.g. ordinary kriging), they may indicate the level of inherent variability within a given site when combined. On this basis, the optimal sampling density, N_{opt}, could be described as:

$$N_{opt} = f\left(\frac{\sum_{i=1}^{j}(w_i CV_i)}{j}\right) \quad \text{Equation 8.1}$$

Here, N_{opt} is a function of the average weighted (w) sum of the coefficients of variation, CV, of *'j'* environmental covariates. Future expansion of this concept should be evaluated over a diverse set of geographical regions. Specifically, how to weight particular coefficients of variation for a given environmental covariate (relative weight between covariates), as well as the provision of the function as some relative metric (relative level of variability between specified areas of comparison).

The optimum sampling investment identified in this work was based on the cost of direct measurement to achieve a certain level of spatial prediction accuracy. However, opportunity exists to improve the accuracy of these spatial predictions with reduced data investment (i.e. directly determine soil characteristics) by augmenting direct measurement with spectral scanning technologies such as Veris (Lund, 1999) or SCANS (Viscarra Rossel et al., 2017). Whilst the measurement accuracy of these technologies is surpassed by direct measurement, their efficiency gains allows for increased spatial data collection, leading to improved spatial predictions, and the possibility of further reducing N_{opt}. Future work must investigate this compromise to determine if

there is viability in these technologies to improve predictions at a reduced cost for the purpose of constraint diagnosis and management.

8.1.2. Amelioration of soil constraints and maintenance of the spatial dataset

Amelioration of soil constraints to realise yield potential requires significant financial investment, perhaps at levels greater than previously considered for many practitioners. For the investigated site, amelioration of sodicity alone required an investment of between 0 to ≈$360/ha for 0–20 cm topsoil treatment, and 0 to ≈$3,300/ha for 0–60 cm profile treatment, depending on the spatial severity. Whilst substantial, the opportunity for increased production due to this investment is equally large. The average yield gap associated with constraints at a regional level has been identified as 0.13 t/ha for sodicity, 0.4 t/ha for acidity and 0.02 t/ha for salinity (Orton et al., 2018). However, this gap is potentially greater when interrogating constraints at a spatially continuous scale as opposed to field average conditions. For the investigated site, the yield gap estimated from this thesis was up to 0.29 t/ha within the 0–20 cm topsoil layer, and 1.4 t/ha within the 0–60 cm profile layer for sodicity alone, resulting in a predicted pay-back time for amelioration investment to be 7 and 10 years. Importantly, constraint amelioration requires long-term strategic economic planning to realise its potential to increase agricultural production. Quite simply, the data taken for soil constraint spatial diagnosis and subsequent management must be thought of as a capital investment, rather than an operational expense.

Importantly, due to the substantial investment within soil sampling alone (≈$28/ha and ≈ $139/ha for 0–20 and 0–60 cm treatment), it is imperative to ensure this data remains relevant as soil amelioration is implemented, which subsequently changes the spatial soil properties where amendment has occurred. Follow-up sampling to investigate the predicted soil response is equivalent to the observed soil response to amendment applied, and if not, refinement of the original dataset should be undertaken to ensure that prediction and models remain relevant for future application. We term this *data maintenance*, which involves follow-up sampling (at a lower density), to track system changes and re-adjust the original dataset such that it remains valid. This approach is similar to machinery maintenance in agricultural production, where a large initial investment is made, but annual maintenance is required to sustain the longevity of the machine. Whilst data maintenance

may be a new concept, we need to ensure these large soil data investments remain applicable through time as amelioration occurs and subsequent chemical and structural changes are made.

We propose the use of permeant monitoring points (PMP), as detailed in McKenzie et al. (2002b) and McKenzie and Dixon (2006), whereby discrete locations are consistently sampled through time to track amendment changes at discrete locations and provide a basis for re-calibration of the original dataset. Future work should focus on the optimal placement of these PMPs such that the spatial variability of structural and chemical change is accurately represented. Appropriate re-sampling times should also be considered, as well as the opportunity to supplement direct measurement with spectroscopic sensing for the purpose of data maintenance.

8.1.3. The requirement for improved interpretability metrics of machine learning models

The phenomena of overfitting and subsequent poor generalisation was observed when machine learning (ML) models were applied to predict yield variability based on soil data. This was apparent when assessing the developed models beyond the traditional method of comparing the difference in R^2 between training and validation. That is, we should be cautious of models which only indicate the goodness of fit by the basic assessment of R^2, especially in highly complex and integrated systems such as agricultural and scale-free natural systems.

In general, models with the smallest difference in R^2 between training and validation exhibited poorer generalization when presented with new information, with the predictions being compared against known soil-yield trends. This highlights the importance of model interpretability, such that the developed ML models can be assessed beyond the R^2 metric. It must be acknowledged by scientists — and understood by the public because these are the people being asked to trust ML black box approaches — that R^2 can often be misleading in assessing generalisation (Alexander et al., 2015). Within the context of soil constraint management, the effects of poor generalisation can be significant, especially where large economic recommendations are made on their basis.

ML interpretability is defined as the "ability to explain or to present in understandable terms to a human" (Doshi-Velez and Kim, 2017). This allows for model outputs to be verified against auxiliary criteria in the assessment of generalisation. Furthermore, assuming model outputs are correct, predicted recommendations are more likely to be accepted by the end-user if the results are explainable in the context of application (Bose and Mahapatra, 2001). From an agronomic

perspective, it is important to provide interpretability metrics for ML models to enable amelioration recommendations to be verified and implemented with confidence surrounding their accuracy. These same metrics may also enable the practitioner to identify when model recommendations are nonsensical, or mechanistically misleading. So, the temporal dimension of the response variable becomes inherently important because a constraint in one year, might not be in another, which could lead to mismanagement, or incorrect diagnosis. A stark example might be where a soil with high ESP has infiltrated water over a very long period of time and in a very dry year is now supporting better growth than previously high performing low ESP areas. The overall yield of the field is extremely low, so relatively it is outperforming in the high ESP areas, but through time it is not. Without the correct domain knowledge of the mechanistic process of dispersion and ESP effect on this, the recommendation for that year would be to increase sodium content of the field. This is clearly a dangerous recommendation, but a practitioner or advisor may implement incorrect advice without domain knowledge. The interpretability of model outputs and design factors that should be considered (e.g. overfitting, generalisation, and output sensibility) renders many ML models as powerless (Vellido et al., 2012), due to the lack of confidence in predictions. Future application of ML within the agricultural domain should therefore on ensuring developed models are more interpretable such that model generalisation can be better assessed and decisions made with improved confidence.

8.1.4. The requirement of increased yield data to overcome temporal model limitations

Two years of yield and weather information was found to be inadequate to train a model to predict yield variation in independent years. This meant the effects of climatic uncertainty could not be accounted for when assessing the yield response due to amelioration, which is required for a more accurate economic analysis of the investment. Whilst estimates could still be achieved using temporal yield averages (i.e. average of 2 seasons), without considering weather information as an independent feature, a requirement exists to build improved models such that this information can be accounted for. The major limitation experienced was insufficient yield data for the investigated site, which is a common problem in the Australian grains industry. Whilst yield monitoring technology has been commercially available for over 2 decades (Pierce et al., 1997), the collected data is often not stored or analysed post-harvest, due to its perceived limited value.

In reality, this yield data by itself does not provide directinsight into system function, unless leveraged using ground-truthed soil measurements to help explain the variability. Conventional soil sampling practices have been inadequate to accurately predict spatial constraints at the resolution required for true precision management. To accurately diagnose soil constraints, by increased soil data collection is necessary along with yield data across multiple years. The implementation of these models subsequently do not allow for the interactions of climate to be accounted for, meaning the ability to forecast future performance is inherently curtailed. Therefore, in order to improve confident in models and investment advice surrounding soil amelioration, yield data and collection is necessary.

8.1.5. Supplementing yield data with biophysical models for temporal predictions

The volume of yield data required to achieve a temporally stable yield prediction model is likely large, and it may not be viable to simply wait for more yield data to become understand the interactions of climate with constraints and yield. An intermediate solution is required. We therefore propose a hybrid biophysical approach to account for temporal variations in weather and the response of yield. Biophysical models such as APSIM (Keating et al., 2003b) are known to be relatively accurate when the soil water retention curve (SWRC) is available for a given location (Archontoulis et al., 2014; Holzworth and Huth, 2011; Whish et al., 2005). Due to the expensive and time consuming nature of measurement, SWRC are rarely available at a spatial scale, and consequently, APSIM predictions are often constrained to point-based applications. However, if SWRC can be measured or predicted at a continuous spatial scale, it would be possible to apply APSIM to augment the detection of a relative yield response across a site, for all years of available weather information.

We propose a regression kriging approach in the estimation of SWRC at a continuous spatial scale, whereby directly measured soil structural and chemical information is used as environmental covariates in the development of an SSPFe. This approach is similar to that of Minasny et al. (2004), Lamorski et al. (2008) and Khlosi et al. (2016). However, the work in this thesis demonstrates that soil dispersion did not correlate well with environmental covariates, therefore would require incorporation of additional dispersion and ESP parameters that are known to be highly influential on soil water dynamics (Bennett et al., 2019; Frenkel et al., 1978).

Obtaining the spatial nature of the SWRC will subsequently allow for the interactions of soil constraints with soil moisture dynamics and crop production. Spatial prediction of plant available water, and the concomitant SWRC, has been widely researched [e.g. (Hedley et al., 2013; Hedley and Yule, 2009; Minasny et al., 1999)]. We advocate that the development of localised calibrations through a mixture of direct sampling and proximal soil sensing is an optimal solution. While there is a clear desire to move away from direct sampling, unreliable prediction of the SWRC suggests that an initial level of direct sampling for localised training data is required to obtain the level of prediction accuracy necessary for variable rate soil management, Future work should detail the required volume of training examples to obtain sufficient accuracy in the prediction of SWRC at the sub-field scale.

This approach does not negate the requirement for yield information, but provides opportunity to explore climate interactions with soil constraints and yield at a greater temporal resolution from which yield data is available. Yield data is, however, still needed in the calibration of the applied APSIM model, with increased yield data resulting in improved calibrations.

8.1.6. Changing the perceived value in Australian agriculture

The perceived value of soil data in Australian agriculture needs to change (Bennett and Cattle, 2013; Bennett and Cattle, 2014). The inherent value of increasing data collection for spatial management of soil constraints has been showcased here. The optimised sampling cost for a variable rate (VR) approach for sodicity amelioration at the investigation site was $28/ha and $139/ha, as opposed to the blanket rate (BR) approach which was limited to $1.38/ha and $2.78/ha, for 0–20 cm and 0–60 cm treatment, respectively. Whilst this difference is substantial, the improved accuracy of the VR gypsum application resulted in: a simulated annual net yield increase of 6.2 t and 26.2 t for the 108 ha site; an increased investment payback time of 5 and 3 years; and, a net positive gain of $27,000 and $104,000 after 20 years (for 0–20 cm and 0–60 cm treatment respectively). This was substantially better than for the BR approach. In this study, the larger initial investment required for a VR approach resulted in the greatest economic gains. Benefits achieved by the soil amelioration far outweigh the initial outlay for the optimal sampling regime of 20 to 50 samples, depending on the soil depth in question. It is clear that data collection has value well beyond that perceived within the current agricultural industries, which suggests a failure of data

consultancies to demonstrate the usefulness of the data collected. Concerted effort is required to overcome this, which is consistent with the findings in the literature (Bennett and Cattle, 2013; Bennett and Cattle, 2014; Lobry de Bruyn, 2019; Lobry de Bruyn and Andrews, 2016; Lobry de Bruyn et al., 2017).

With the increased initial cost of sampling for a VR approach to soil amelioration, but potential long-term economic benefits of improved application, soil sampling investment needs to be viewed as a capital investment, not an operational expense. The greatest economic gains from improved application are obtained from long-term yield increases, as opposed to savings due to over application of resource which is realised in the first year. Therefore, it cannot be expected to recover the cost of sampling within a short time period, and instead, a long-term strategic approach to sampling investment needs to be made. This highlights the requirement for sound economics to be dovetailed with constraint management programs. Furthermore, identifying data collection for soil constraints as a capital investment is not only important for growers and practitioners, but financiers as well, as they need to be inherently aware of the cost-benefit relationship with soil sampling and agronomic benefit, in order to give growers the flexibility to make investments for sustained profitability within their farming business.

It is important to note that the VR approach discussed here is not based on zone management (current VR standard). It is instead based on a DSM approach, whereby soil properties are mapped on a continuous scale and are not limited to hard boundaries. Whilst zone management offers improved accuracies over a BR approach, it remains severely limited as a VR approach to accurately identify and manage soil trends to the level required for true precision management. As found here, a better approach is to increase sampling investment to achieve continuous mapping of soil properties. Zone management should not be considered a true VR approach for soil amelioration, and the language used by practitioners and advisors needs to shift to reflect this.

8.1.7. From local to universal calibration

The work presented here has developed a locally calibrated model to explore the interactions between soil constraints and yield, using an extensive dataset collected for a single site. The model optimsed investment strategies pertaining to sampling densities and soil

amelioration, although it is unlikely to produce accurate predictions at unvisited sites, as the model is constrained by the range of soil characteristics presented at the training site. In fact, the model was tested using 30 sites external to the localised calibration and it was found to perform extremely poorly (results not presented), primarily because it was presented with data outside of the training range. Furthermore, if the training site does exhibit sufficient soil observations with unconstrained properties, the model cannot accurately predict the yield response due to amelioration, therefore providing conservative estimates of the likely benefit; i.e. where true unconstrained data can be incorporated, then the model better fits the true yield potential of a given site.

If the process of increased data collection is repeated for a number of sites across a region (i.e. it becomes an industry norm to provide capital investment in soil data), opportunity exists to merge this data in an attempt to build a universally calibrated model to explore the effects of soil constraints on yield. This would expose the model to a wider variety of soil constraint and weather interactions therefore offering improved capabilities of prediction at unvisited sites. Additionally, the increased data would improve crop response predictions within individual sites, in terms of the true yield potential. We refer to this a *data collaboration*, where individual datasets are regionally linked to provide a more powerful outcome than the datasets could achieve individually. Such *data collaboration* moves further towards the ability to apply more advanced NLML, so must also be undertaken with respect to the considerations around interpretability. In terms of *data collaboration*, there are also privacy, access, and legal considerations associated with the release of data. However, anecdotally, *data collaboratives* are establishing within Australia, and as a collaborative will have a greater potential and power to deal with privacy, access, and legal considerations.

Achieving universal calibration through *data collaboration* requires a level of standardization for data collection to ensure results between sites are interoperable, comparable and reusable for purpose. This will involve accurate recording keeping, incorporation of management information and a set of required soil characteristics to measure as well as a set of laboratory standards coupled with these. This should remain a focus of precision agriculture, where regional collaboration is used to provide improved agronomic benefit for the region as whole.

8.1.8. Towards an integrated farming systems model

Within the farming system, there is a plethora of information available that provides opportunities for improved decision making, if collected and processed to provide useful insight into the complexities of the system. Inhibiting our ability to achieve this is the diverse nature of data within the farming system, which may be quantitative, qualitative, or a mixture of both. Furthermore, a multitude of data sources provide difficulties in capturing and utilising this information, which can be categorised into *process-mediated* (PM), *machine-generated* (MG) and *human sourced* (HS) data (Bennett, 2015; Devlin, 2012). PM data is that which generated in the running of the agricultural enterprise including agricultural inputs and business transactions (e.g. soil tests, nutrient rates and spatial application etc.), while MG data is collected in an automated sense, either on-farm or externally, through sensors and stored on databases (e.g. climate data, capacitance probe information etc.). The more difficult data to handle is HS, which is generated in the carrying out of everyday life and can be as simple as discussion (face to face, or online), management based constraints to implementation of best management practices, and patterns in the way that information is searched/consumed (e.g. click patterns on the internet). Harnessing this full gambit of information provides the capability to provide more informed recommendations, and can lead to the realisation of action that might previously not have been considered (Bennett, 2015; Kelly et al., 2017).

Presented in this thesis is the development and application of empirical, biophysical and probabilistic approaches that utilise each of these data sources to provide meaningful information for implementation in the decision making process. There is a clear requirement to develop an integrated farming systems framework (IFSF) that combines these data handling and analytic approaches to provide more accurate agronomic recommendations across a wider range of system variables. An IFSF would combine the various PM, MG and HS data through a range of avenues influencing the operational and economic on-farm models by integrating these data streams, and concomitant considerations, prior to the subsequent decision point. This decision point now being more informed.

We propose the future development of an IFSF, based on the conceptual framework presented in Figure 8.1, to provide a basis for decision making. Within the framework, inference

can be provided by application of four main modelling approaches that aim to incorporate varied data sources, namely: i) management probabilistic models; ii) biophysical models; iii) process probabilistic modes; and, iv) empirical models. Management probabilistic models aim to handle human-sourced data and biophysical model outputs to translate information in a meaningful form for empirical modelling. The intention is that this data should feed into categorical weightings within empirical models to incorporate management nuance within the temporal aspects of the empirical approach. An example of this may include describing a crop planting date as 'early, 'average', or 'late' as opposed to an exact date, which is meaningless to an empirical model, but is highly influential on the outcome of the system. Where the HS management information cannot directly be incorporated into the empirical modelling process to influence the outcomes, it can be directly imposed into the economic assessment of model outcomes.

Process probabilistic models utilise external, MG and PM data with biophysical outputs to account for system variables that cannot be measured or sensed quantitatively (e.g. soil compaction). Managing these variables may result in an economic gain, however, they cannot be assessed using empirical approaches due to their qualitative nature. Examples of this may include erosion or compaction, which cannot easily be measured quantitatively, but effect soil function and subsequently economic performance of the farming system. The use of risk based approaches through probabilistic modelling consequently allow these data to be included within the decision process, as was demonstrated in Chapter 7.

Empirical models within the framework capture quantitative on-farm and external MG and PM data, or quantitative outputs of other models. Provided the data volume is sufficient, these approaches can be powerful in modelling agronomic processes, however, require purely quantitative information. Biophysical models are currently an integral component of the IFSF to model biological processes. However, as data volume increases, empirical approaches may become more powerful at modelling these processes and subsequently negate the need for biophysical models.

Within the IFSF, all model outputs culminate into an economic model, as economic metrics are used as the initial basis for on-farm decisions; i.e. what is the economic ramification of the intervention and does it fit the production requirement of the enterprise? This does not mean that natural and social capital aren't considered; indeed this is incorporated at the decision point

through the *social, faming, and farming organisation norms* that define non-economic values of the farming organisation known to shape the decision process. Importantly, the model based aspects of the IFSF terminates at a recommendation node, where the developed agronomic recommendations are subsequently synthesised according to these farming organisation norms, as well as with trusted consultants. The presented approach does not aim to negate consultants or advisors, but instead equip the farming organisation, including advisors/consultants, with tools to facilitate improved decision making processes on farm.

The presented IFSF conceptualises an approach to utilise various data sources to aid in on-farm decision making process. As farming systems evolve, this framework can adapt to represent new variables, processes and data source. We therefore see this framework as a dynamic roadmap to on-farm decision making.

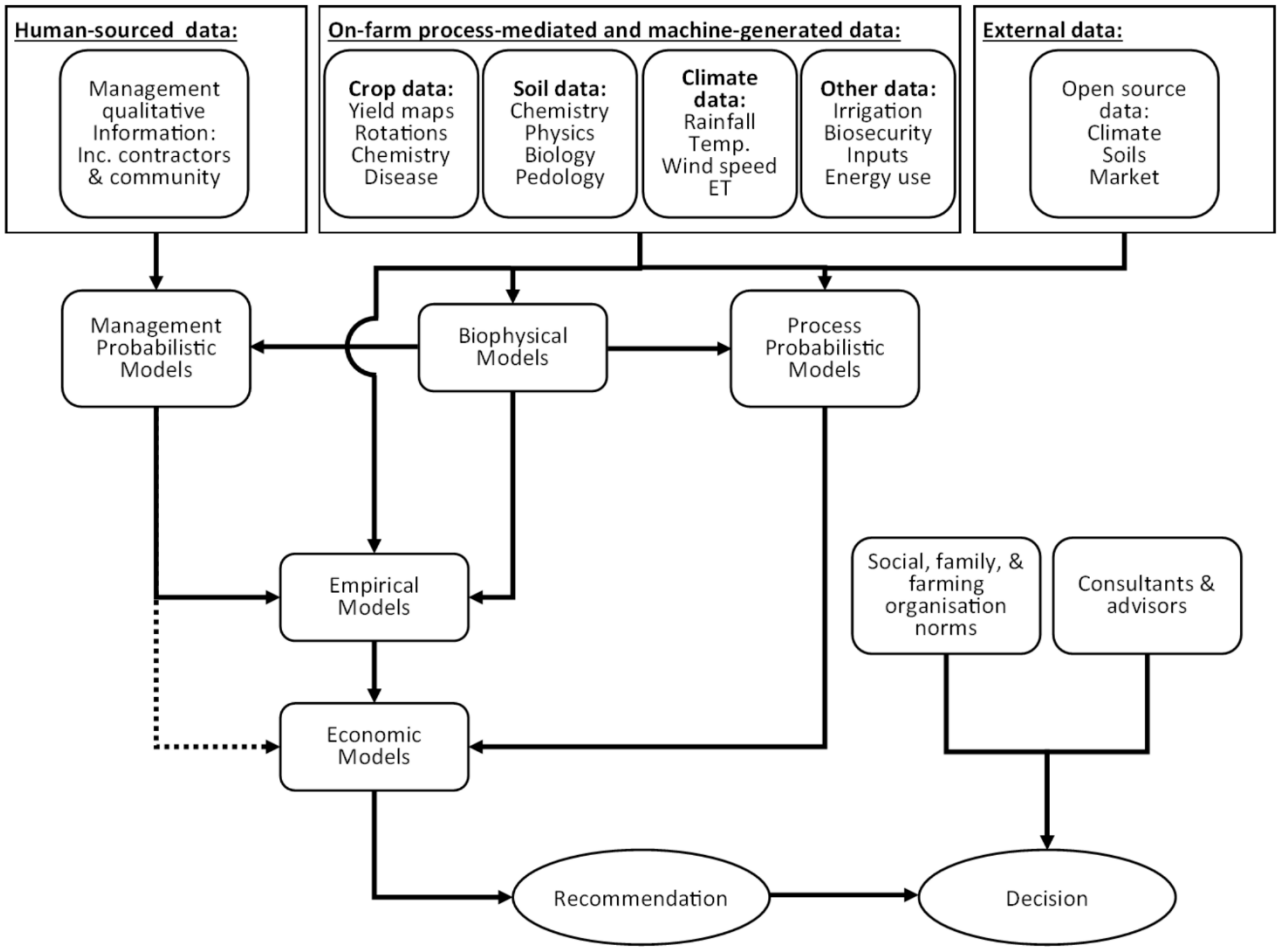


Figure 8.1. Proposed integrated farming system framework (IFSF) to merge multiple data sources in a hybrid modeling approach to inform on-farm decision making. Arrows represent the directional flow of information through the framework. ET within the climate data category represents evapotranspiration

8.2. General conclusions

From the topics and themes investigated in this thesis, the following general conclusions are drawn. With regard to the agronomic consequences of a data minimal approach to soil sampling, some conclusions are:

- Sampling design becomes less important, whereby splitting the field into area equivalent strata and taking a random sample from within this resulted in sufficient description of variability, often superior to use of more advanced geostatistical approaches. The sampling density then becomes the point of consideration.

- Increasing sampling density greatly improves the accuracy of agronomic recommendations for soil amelioration. This is due to a more spatially accurate ameliorant recommendation that results in reduced over-application of resources.

- For sodicity amelioration, the economically optimal sampling density for 0–20 cm and 0–60 cm treatment is within the vicinity of 0.2 cores/ha (1 core per 5 ha) and 0.5 cores/ha (1 core per 2 ha), depending on the inherent variability presented at the sampling site.

- A variable rate approach to soil amelioration is economically superior over blanket-rate application, despite a larger sampling investment being required. A VR approach (based on the continuous mapping of soil properties, not zone management) should remain the focus of amelioration advice.

- Soil sampling should be seen as a capital investment, not an operational cost within the farming business. The suggested sampling investments require an increased upfront cost in comparison to traditional approaches, however, the long term benefits as the result of improved ameliorant recommendation and subsequent yield response is substantial.

- The majority of yield response due to amelioration is achieved from small discrete areas. A requirement therefore exists to ensure the scale of spatial management in matched to the scale of these discrete areas that remain highly economically significant, in terms of their yield potential.

With regard to the accuracy of digital soil mapping approaches to spatially predict soil properties, some conclusions are:

- The widely adopted bulked transect sampling method to obtain field average conditions is highly inaccurate and presents large economic concern for soil amelioration advice on its basis. Furthermore, the results obtained from this method are highly sensitive to the random initialisation of the sampling transect. This method should not be used for soil amelioration advice

- A regression kriging approach to DSM (via application of SSPFe) is dependent on the strengths of correlation between the environmental covariates and the target variable. If poor correlation exists, geostatistical methods such as ordinary kriging may be more beneficial. This was observed here.

- Increasing sampling density greatly improves the spatial prediction accuracy of regression kriging and ordinary kriging, the large majority of which was observed up to 1 sample/ha. Increasing sampling density did not greatly improve the accuracy of bulked transect sampling or sampling for zone management

With regard to the merit of linear and non-linear machine learning approaches to reveal key site-specific soil-crop interactions, some conclusions are:

- The developed Cubist model provided superior results and confidence in predicting single-season yield variability in comparison to mixed-linear regression, artificial neural networks and support vector machines. Furthermore, Cubist provides a more interpretable model.

- Increased yield data is required to strengthen the development of yield variability prediction by allowing for climatic uncertainty to be incorporated. Using the available 3 years of wheat yield data, single-year models could only be developed

- Better interpretability metrics are required to assess the performance and generalisation of machine learning models within the context of agriculture. Relying on the R^2 metric may provide misleading confidence on the capabilities of the model, leading to inaccurate agronomic recommendation advice.

- Data dimensionality reduction as a pre-processing technique for high-dimensional spatial is pertinent in aiding model convergence and improving generalisation. This can be achieved using principal component analysis (PCA)

- Ordinary kriging as a data augmentation technique is highly beneficial to artificially increase the training data size to aid in improved generalisation. Without augmentation, spatial datasets are of an insufficient size to train machine learning models

With regard to the ability to capture and integrate qualitative management information to inform decision making, some conclusions are:

- A probability based approach allows the inclusion of qualitative data, providing output that can be utilised to make management decisions based upon risk. The compaction example applied in this work could easily be extrapolated to other difficult constructs that affect biophysical model output.

- Management nuance is too varied to collect at a data resolution sufficient to obtain complete feature sets for inclusion in quantitative modelling. Therefore, the use of BBN integrated with the other approaches used in this thesis will allow a more complete farming systems model that has capability to provide meaningful decision making advice.

8.3. Future research directions

Various directions for future research have been identified throughout the numerous chapters of this thesis. These are outlined below.

To appropriately manage the spatial nature of soil constraints, it is imperative to ensure soil data collection is of an adequate density to accurately describe this variability and subsequently provide targeted soil amelioration advice. The economically optimal sampling density (N_{opt}) required to provide this advice for a single site is presented in this thesis, however, further work is required to extend these findings to inform optimal sampling densities at unvisited locations, where no prior information pertaining to the target variable is known. The optimal sampling density is dependent on the variability presented at a given site, and as such, we propose an integrated approach whereby environmental covariates are combined to provide a proxy estimate of this variability. This would allow for the estimation of N_{opt} when only these covariates were available. Environmental covariates may include yield data (of multiple seasons), remotely sensed data (e.g. multispectral or hyperspectral satellite or drone data) and proximally sensed data (electromagnetic induction, gamma-ray spectrometry). Achieving this will require additional sites to be investigated across different agro-ecosystems in a similar fashion to that presented here, in order to calibrate this proxy measure to inform optimum sampling densities for constraint management using only environmental covariates. This approach may also incorporate qualitative information pertaining to farmer knowledge of spatial variability. If successful, this approach would effectively identify N_{opt} using spatial data layers that were either already available, or could be cheaply acquired, therefore providing an economical approach for site characterisation.

The investment required to obtain sufficient samples for a localised calibration to spatially diagnose constraints and their effects on yield is substantial. Similar to any other capital on-farm investment, a level of maintenance is required to ensure the purchased item (in this case a soil dataset) remains relevant and applicable for future application. This will require *data maintenance*, where follow-up sampling is undertaken to track soil trends and calibrate the original dataset to remain relevant. A logical path forward to achieve this is by use of permanent monitoring points (PMPs) that are frequently sampled. These points should be located such that they remain representative of site, as calculated using the original directly measured dataset. The frequency of sampling also needs to be considered in accordance to the likely speed of soil change (i.e. re-

sampling periods of 1–5 years may suffice for structural properties, but <1 year may be necessary for specific nutrient properties). Therefore, the frequency of sampling will likely be different in accordance to the target variable. Proximal sensing, remote sensing and yield information may provide useful data layers in augmenting these PMPs to extrapolate the measured trends to all areas across the site. Furthermore, these data layers provide a means to track independent spatiotemporal trends that may diverge from that observed when the original dataset was collected. This may inform the need for increased PMPs at new locations.

Investigations presented in this thesis were limited by the availability of yield information in the development of a temporally stable yield prediction model to account for the uncertainties of climate interactions with the spatial variability of yield. A potential way forward to overcome this involves a hybrid approach whereby biophysical models, such as APSIM, are employed at a sub-field spatial scale. This requires the spatial prediction of soil-water retention curves, which we envisage will involve the use of direct sampling and proximal soil sensing. Due to the time consuming and costly nature of SWRC analysis, future work should detail the minimum directly measured dataset for a practically useful spatial prediction of this specific property — practical usefulness being defined as the ability to employ on-farm and provide a level of efficiency beyond the current system, rather than being accurate to the n^{th} degree. Furthermore, there is immense opportunity to supplement direct measurement of SWRC with PSS to aid in the spatial interpolation of the SWRC. Identifying the protocols for SWRC local calibration and spatial interpretation will require relatively intensive sampling across a diverse geography to understand the spatial trends in SWRC and their effects of biophysical model yield predictions. As a first step, we propose re-visiting the site investigated in this work to obtain SWRC data at equivalent locations to the already obtained soil structural and chemical data. Subsequently, more investigation sites will be required to extend the findings to new unvisited locations.

When attempting to develop site-specific variogram models for spatial prediction, future sampling designs should also consider the collection of short-range samples independently of the main dataset, to improve estimates of the variogram parameters at shorter lags, as described in (Webster and Lark, 2012). Whilst not essential, these additional samples may help the overall spatial prediction, or at the very least, identify the short-range errors of the variogram model.

Furture work should consider the cost of additional data collection for this purpose in comparison to the relative improvements of spatial predictions.(Webster and Lark, 2012)

Local calibration datasets, such as that developed in this body of work, are limited by the ability to make predictions at independent locations. This is due to the training data being bounded by the range of conditions observed within a single site. Universal calibration of the developed models will require additional soil data collected across a diverse geographic region to increase the range of training observations. Through *data collaboration*, these models will become increasingly powerful in providing soil-crop inference. In the merging of this data to provide inference at a new site, a distance-weighted metric may be required to bias both the geographically closest datasets, as well as the pedometrically closest datasets (Huang et al., 2018). Furthermore, a management metric may also be required to bias against datasets obtained from farming systems under similar management (e.g. CTF).

8.4. Entire reference list

Aalders, I., 2008. Modeling land-use decision behavior with Bayesian belief networks. Ecology and Society 13(1).

ABARES, 2019. Weekly Australian Climate, Water and Agricultural Update. Department of Agriculture and Water Resouces.

Abbaspour, K., Schulin, R., van Genuchten, M.T., Schläppi, E., 1998. An alternative to cokriging for situations with small sample sizes. Mathematical geology 30(3), 259-274.

Abbott, T., McKenzie, D., 1986. Improving soil structure with gypsum. Department of Agriculture. New South Wales, Agfact AC 10.

Abu-Hamdeh, N.H., 2003. Compaction and subsoiling effects on corn growth and soil bulk density. Soil Science Society of America Journal 67(4), 1213-1219.

Adamchuk, V.I., Hummel, J., Morgan, M., Upadhyaya, S., 2004. On-the-go soil sensors for precision agriculture. Computers and electronics in agriculture 44(1), 71-91.

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environmental Modelling & Software 26(12), 1376-1388.

Ahrens, R.J., 2008. Digital soil mapping with limited data. Springer Science & Business Media.

Alaoui, A., Lipiec, J., Gerke, H., 2011. A review of the changes in the soil pore system due to soil deformation: A hydrodynamic perspective. Soil and Tillage Research 115, 1-15.

Alexander, D., Tropsha, A., Winkler, D.A., 2015. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. Journal of chemical information and modeling 55(7), 1316-1322.

Alzubaidi, A., Webster, G., 1982. Effect of tillage in combination with chemical amendments on reclamation of a solonetzic soil. Canadian Journal of Soil Science 62(4), 641-649.

Amini, M., Abbaspour, K.C., Khademi, H., Fathianpour, N., Afyuni, M., Schulin, R., 2005. Neural network models to predict cation exchange capacity in arid regions of Iran. European Journal of Soil Science 56(4), 551-559.

Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure, ACM Sigmod record. ACM, pp. 49-60.

Antille, D.L., Bennett, J.M., Jensen, T.A., 2016a. Soil compaction and controlled traffic considerations in Australian cotton-farming systems. Crop and Pasture Science 67(1), 1-28.

Antille, D.L., Bennett, J.M., Jensen, T.A., Roberton, S.D., 2016b. The influence of tyre inflation pressure on soil compaction caused by the John Deere 7760. An impact assessment framework for harvesting technologies in cotton: Management considerations forthe John Deere 7760 National Centre for Engineering in Agriculture Publication 1004960/16/1, USQ, Toowoomba.

Archontoulis, S.V., Miguez, F.E., Moore, K.J., 2014. Evaluating APSIM maize, soil water, soil nitrogen, manure, and soil temperature modules in the Midwestern United States. Agronomy Journal 106(3), 1025-1040.

Armstrong, L.J., Diepeveen, D., Maddern, R., 2007. The application of data mining techniques to characterize agricultural soil profiles, pp. 85-100.

Arslan, S., Colvin, T.S., 2002. Grain yield mapping: Yield sensing, yield reconstruction, and errors. Precision Agriculture 3(2), 135-154.

Arthur, E., Moldrup, P., Schjønning, P., de Jonge, L.W., 2013. Water retention, gas transport, and pore network complexity during short-term regeneration of soil structure. Soil Science Society of America Journal 77(6), 1965-1976.

Atzberger, C., 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote Sensing 5(2), 949-981.

Australian Bureau of Statistics, 2015. Agricultural Commodities, Australia. cat. no. 7121.0.

Aylmore, L., Karim, M., Quirk, J., 1971. Dissolution of gypsum, monocalcium phosphate, and superphosphate fertilizers in relation to particle size and porous structure. Soil Research 9(1), 21-32.

Baharom, S.N.A., Shibusawa, S., Kodaira, M., Kanda, R., 2015. Multiple-depth mapping of soil properties using a visible and near infrared real-time soil sensor for a paddy field. Engineering in agriculture, environment and food 8(1), 13-17.

Bailey, A., Raper, R., Johnson, C., Burt, E., 1995. An integrated approach to soil compaction prediction. Journal of Agricultural Engineering Research 61(2), 73-80.

Baker, L., Ellison, D., 2008. Optimisation of pedotransfer functions using an artificial neural network ensemble method. Geoderma 144(1-2), 212-224.

Baldi, P., Hornik, K., 1989. Neural networks and principal component analysis: Learning from examples without local minima. Neural networks 2(1), 53-58.

Ballabio, C., 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151(3-4), 338-350.

Bargoti, S., Underwood, J., 2017. Deep fruit detection in orchards, 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 3626-3633.

Barnes, E., Clarke, T., Richards, S., Colaizzi, P., Haberland, J., Kostrzewski, M., Waller, P., Choi, C., Riley, E., Thompson, T., 2000. Coincident detection of crop water stress, nitrogen status and canopy

density using ground based multispectral data, Proceedings of the Fifth International Conference on Precision Agriculture, Bloomington, MN, USA.

Bartimote, T., Quigley, R., Bennett, J.M., Hall, J., Brodrick, R., Tan, D.K., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: II. Economic and physiological analysis. Soil and Tillage Research 168, 133-142.

Bashari, H., Smith, C., Bosch, O., 2008. Developing decision support tools for rangeland management by combining state and transition models and Bayesian belief networks. Agricultural Systems 99(1), 23-34.

Behmann, J., Mahlein, A.-K., Rumpf, T., Römer, C., Plümer, L., 2015. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. Precision Agriculture 16(3), 239-260.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Journal of plant nutrition and soil science 168(1), 21-33.

Bennett, J.M., 2015. Agricultural big data: utilisation to discover the unknown and instigate practice change. Farm Policy Journal 12(1), 43-50.

Bennett, J.M., 2019. Soil Security for Australia. Soil Systems.

Bennett, J.M., Cattle, S., 2013. Adoption of soil health improvement strategies by Australian farmers: I. Attitudes, management and extension implications. The Journal of Agricultural Education and Extension 19(4), 407-426.

Bennett, J.M., Cattle, S., 2014. Adoption of soil health improvement strategies by Australian farmers: II. Impediments and incentives. The Journal of Agricultural Education and Extension 20(1), 107-131.

Bennett, J.M., Cattle, S., Singh, B., 2015a. The efficacy of lime, gypsum and their combination to ameliorate sodicity in irrigated cropping soils in the Lachlan Valley of New South Wales. Arid Land Research and Management 29(1), 17-40.

Bennett, J.M., Marchuk, A., Raine, S., Dalzell, S., Macfarlane, D., 2016. Managing land application of coal seam water: A field study of land amendment irrigation using saline-sodic and alkaline water on a Red Vertisol. Journal of environmental management 184, 178-185.

Bennett, J.M., Roberton, S., Marchuk, S., Woodhouse, N., Antille, D., Jensen, T., Keller, T., 2019. The soil structural cost of traffic from heavy machinery in Vertisols. Soil and Tillage Research 185, 85-93.

Bennett, J.M., Roberton, S.D., Jensen, T.A., Antille, D.L., Hall, J., 2017. A comparative study of conventional and controlled traffic in irrigated cotton: I. Heavy machinery impact on the soil resource. Soil and Tillage Research 168, 143-154.

Bennett, J.M., Woodhouse, N.P., Keller, T., Jensen, T.A., Antille, D.L., 2015b. Advances in cotton harvesting technology: a review and implications for the John Deere round baler cotton picker. Journal of Cotton Science 19(2), 225-249.

Bentley, M.L., Mote, T.L., Thebpanya, P., 2002. Using Landsat to identify thunderstorm damage in agricultural regions. Bulletin of the American Meteorological Society 83(3), 363-376.

Besalatpour, A., Hajabbasi, M., Ayoubi, S., Afyuni, M., Jalalian, A., Schulin, R., 2012. Soil shear strength prediction using intelligent systems: artificial neural networks and an adaptive neuro-fuzzy inference system. Soil science and plant nutrition 58(2), 149-160.

Bezdek, J.C., 1975. Mathematical models for systematics and taxonomy, Proceedings of eighth international conference on numerical taxonomy, pp. 143-166.

Bi, C., Chen, G., 2011. Bayesian Networks Modeling for Crop Diseases. In: D. Li, Y. Liu, Y. Chen (Eds.), Computer and Computing Technologies in Agriculture IV: 4th IFIP TC 12 Conference, CCTA 2010, Nanchang, China, October 22-25, 2010, Selected Papers, Part I. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 312-320.

Birant, D., Kut, A., 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data & Knowledge Engineering 60(1), 208-221.

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Bishop, T., Horta, A., Karunaratne, S., 2015. Validation of digital soil maps at different spatial supports. Geoderma 241, 238-249.

Bishop, T., McBratney, A., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103(1-2), 149-160.

Bogaert, P., Russo, D., 1999. Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. Water Resources Research 35(4), 1275-1289.

Boroughs, P., 1986. Principles of Geographic Information Systems for Land Resource Assessment. Beckett, PHT.

Bose, I., Mahapatra, R.K., 2001. Business data mining—a machine learning perspective. Information & management 39(3), 211-225.

Botta, G.F., Pozzolo, O., Bomben, M., Rosatto, H., Rivero, D., Ressia, M., Tourn, M., Soza, E., Vazquez, J., 2007. Traffic alternatives for harvesting soybean (Glycine max L.): effect on yields and soil under a direct sowing system. Soil and Tillage Research 96(1-2), 145-154.

Bowman, K., 2008. Economic and environmental analysis of converting to controlled traffic farming, Proceedings 6th Australian Controlled Traffic Farming Conference, pp. 12-14.

Boydell, B., McBratney, A., 2002. Identifying potential within-field management zones from cotton-yield estimates. Precision agriculture 3(1), 9-23.

Bragato, G., 2004. Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain. Geoderma 118(1-2), 1-16.

Braunack, M., Johnston, D., 2014. Changes in soil cone resistance due to cotton picker traffic during harvest on Australian cotton soils. Soil and Tillage Research 140, 29-39.

Briscoe, E., Feldman, J., 2011. Conceptual complexity and the bias/variance tradeoff. Cognition 118(1), 2-16.

Brus, D., De Gruijter, J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80(1-2), 1-44.

Brus, D., De Gruijter, J., Van Groenigen, J., 2004. Designing purposive and random spatial coverage samples by the k-means clustering algorithm, Global Workshop on Digital Soil Mapping, Montpellier, France.

Brus, D., Kempen, B., Heuvelink, G., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62(3), 394-407.

Brus, D.J., Heuvelink, G.B., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138(1-2), 86-95.

Budayan, C., Dikmen, I., Birgonul, M.T., 2009. Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. Expert Systems with Applications 36(9), 11772-11781.

Buhmann, M.D., 2003. Radial basis functions: theory and implementations, 12. Cambridge university press.

Bünemann, E.K., Bongiorno, G., Bai, Z., Creamer, R.E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T.W., Mäder, P., 2018. Soil quality–A critical review. Soil Biology and Biochemistry 120, 105-125.

Burgess, T., Webster, R., 1980a. Optimal interpolation and isarithmic mapping of soil properties: I. The semivariogram and punctual kriging. Journal of soil science 31(2), 315-331.

Burgess, T., Webster, R., 1980b. Optimal interpolation and isarithmic mapping of soil properties: II block kriging. Journal of Soil Science 31(2), 333-341.

Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Soria-Olivas, E., Martín-Guerrero, J.D., Moreno, J., 2003. Support vector machines for crop classification using hyperspectral data, Iberian Conference on Pattern Recognition and Image Analysis. Springer, pp. 134-141.

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007. Digital soil assessments: Beyond DSM. Geoderma 142(1-2), 69-79.

Chamen, T., Alakukku, L., Pires, S., Sommer, C., Spoor, G., Tijink, F., Weisskopf, P., 2003. Prevention strategies for field traffic-induced subsoil compaction: a review: Part 2. Equipment and field practices. Soil and Tillage Research 73(1), 161-174.

Chamen, W., Longstaff, D., 1995. Traffic and tillage effects on soil conditions and crop growth on a swelling clay soil. Soil Use and Management 11(4), 168-176.

Chan, A.B., Vasconcelos, N., Lanckriet, G.R., 2007. Direct convex relaxations of sparse SVM, Proceedings of the 24th international conference on Machine learning. ACM, pp. 145-153.

Chan, K., Oates, A., Swan, A., Hayes, R., Dear, B., Peoples, M., 2006. Agronomic consequences of tractor wheel compaction on a clay soil. Soil and Tillage Research 89(1), 13-21.

Chang, D.-H., Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural network. Remote Sensing of Environment 74(3), 534-544.

Chawla, V., Naik, H.S., Akintayo, A., Hayes, D., Schnable, P., Ganapathysubramanian, B., Sarkar, S., 2016. A Bayesian Network approach to County-Level Corn Yield Prediction using historical data and expert knowledge. arXiv preprint arXiv:1608.05127.

Chiang, M.M.-T., Mirkin, B., 2010. Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. Journal of classification 27(1), 3-40.

Chuang, C.-C., Su, S.-F., Jeng, J.-T., Hsiao, C.-C., 2002. Robust support vector regression networks for function approximation with outliers. IEEE Transactions on Neural Networks 13(6), 1322-1330.

Cockx, L., Van Meirvenne, M., Vancoillie, F., Verbeke, L., Simpson, D., Saey, T., 2010. A Neural Network Approach to Topsoil Clay Prediction Using an EMI-Based Soil Sensor, Proximal Soil Sensing. Springer, pp. 245-254.

Cressie, N., 1985. Fitting variogram models by weighted least squares. Journal of the International Association for Mathematical Geology 17(5), 563-586.

Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Dai, X., Huo, Z., Wang, H., 2011. Simulation for response of crop yield to soil moisture and salinity with artificial neural network. Field Crops Research 121(3), 441-449.

Dang, A., Bennett, J.M., Marchuk, A., Biggs, A., Raine, S., 2018. Quantifying the aggregation-dispersion boundary condition in terms of saturated hydraulic conductivity reduction and the threshold electrolyte concentration. Agricultural water management 203, 172-178.

Dang, Y., Dalal, R., Mayer, D., McDonald, M., Routley, R., Schwenke, G., Buck, S., Daniells, I., Singh, D., Manning, W., 2008. High subsoil chloride concentrations reduce soil water extraction and crop yield on Vertosols in north-eastern Australia. Australian Journal of Agricultural Research 59(4), 321-330.

Dang, Y., Dalal, R., Routley, R., Schwenke, G., Daniells, I., 2006. Subsoil constraints to grain production in the cropping soils of the north-eastern region of Australia: an overview. Australian Journal of Experimental Agriculture 46(1), 19-35.

Dang, Y., Moody, P., 2016. Quantifying the costs of soil constraints to Australian agriculture: a case study of wheat in north-eastern Australia. Soil Research 54(6), 700-707.

Dang, Y.P., Dalal, R.C., Buck, S., Harms, B., Kelly, R., Hochman, Z., Schwenke, G.D., Biggs, A., Ferguson, N., Norrish, S., 2010. Diagnosis, extent, impacts, and management of subsoil constraints in the northern grains cropping region of Australia. Soil Research 48(2), 105-119.

DasGupta, B., Schnitger, G., 1993. The power of approximating: a comparison of activation functions, Advances in neural information processing systems, pp. 615-622.

Davidson, J., Quirk, J., 1961. The influence of dissolved gypsum on pasture establishment on irrigated sodic clays. Australian Journal of Agricultural Research 12(1), 100-110.

De Gruijter, J., Brus, D.J., Bierkens, M.F., Knotters, M., 2006. Sampling for natural resource monitoring. Springer Science & Business Media.

De Gruijter, J., McBratney, A., 1988. A modified fuzzy k-means method for predictive classification.

De Gruijter, J.J., 1977. Numerical classification of soils and its application in survey. Pudoc.

Deere and Company, 2018a. Product brochure: 4 Series Self-Propelled Sprayers.

Deere and Company, 2018b. Product brochure: 8R/8RT Series.

Deere and Company, 2018c. Product brochure: Combines.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., McLoone, S., 2013. Principal component analysis on spatial data: an overview. Annals of the Association of American Geographers 103(1), 106-128.

Devlin, B., 2012. The Big Data Zoo—Taming the Beasts: The need for an integrated platform for enterprise information. Cape Town: 9sight Consulting.

Dlugoß, V., Fiener, P., Schneider, K., 2010. Layer-specific analysis and spatial prediction of soil organic carbon using terrain attributes and erosion modeling. Soil Science Society of America Journal 74(3), 922-935.

Doerge, T., 1999. Defining management zones for precision farming. Crop Insights 8(21), 1-5.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Doyle, R., Habraken, F., 1993. The distribution of sodic soils in Tasmania. Soil Research 31(6), 931-947.

Drummond, S., Joshi, A., Sudduth, K.A., 1998. Application of neural networks: precision farming, Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on. IEEE, pp. 211-215.

Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J., Kitchen, N.R., 2003. Statistical and Neural Methods for Site–Specific Yield Prediction. Transactions of the ASAE 46(1), 5.

Drury, B., Valverde-Rebaza, J., Moura, M.-F., de Andrade Lopes, A., 2017. A survey of the applications of Bayesian networks in agriculture. Engineering Applications of Artificial Intelligence 65, 29-42.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.

Eitrich, T., Lang, B., 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. Journal of Computational and Applied Mathematics 196(2), 425-436.

Ennett, C.M., Frize, M., Walker, C.R., 2001. Influence of missing values on artificial neural network performance, Medinfo, pp. 449-453.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, Kdd, pp. 226-231.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1998. Clustering for mining in large spatial databases. KI 12(1), 18-24.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S., Bouras, A., 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing 2(3), 267-279.

Faithpraise, F., Birch, P., Young, R., Obu, J., Faithpraise, B., Chatwin, C., 2013. Automatic plant pest detection and recognition using k-means clustering algorithm and correspondence filters. International Journal of Advanced Biotechnology and Research 4(2), 189-199.

Farmani, R., Henriksen, H.J., Savic, D., 2009. An evolutionary Bayesian belief network methodology for optimum management of groundwater contamination. Environmental Modelling & Software 24(3), 303-310.

Fawzi, A., Samulowitz, H., Turaga, D., Frossard, P., 2016. Adaptive data augmentation for image classification, 2016 IEEE International Conference on Image Processing (ICIP). Ieee, pp. 3688-3692.

Fiener, P., Auerswald, K., 2009. Spatial variability of rainfall on a sub ‐ kilometre scale. Earth Surface Processes and Landforms 34(6), 848-859.

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

Florinsky, I.V., Eilers, R.G., Manning, G., Fuller, L., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling & Software 17(3), 295-311.

Ford, G., Martin, J., Rengasamy, P., Boucher, S., Ellington, A., 1993. Soil sodicity in Victoria. Soil Research 31(6), 869-909.

Frenkel, H., Goertzen, J., Rhoades, J., 1978. Effects of clay type and content, exchangeable sodium percentage, and electrolyte concentration on clay dispersion and soil hydraulic conductivity 1. Soil Science Society of America Journal 42(1), 32-39.

Fu, Q., Wang, Z., Jiang, Q., 2010. Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO. Mathematical and computer modelling 51(11-12), 1299-1305.

Gee, G., Bauder, J., 1986. Particle-size analysis. In 'Methods of soil analysis. Part 1. Physical and mineralogical methods'.(Ed. A Klute) pp. 383–411. Soil Science Society of America: Madison, WI, USA.

Genc, H., Genc, L., Turhan, H., Smith, S., Nation, J., 2008. Vegetation indices as indicators of damage by the sunn pest (Hemiptera: Scutelleridae) to field grown wheat. African Journal of Biotechnology 7(2).

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines. JAWRA Journal of the American Water Resources Association 42(4), 1033-1046.

Gnatowski, T., Szatyłowicz, J., Brandyk, T., Kechavarzi, C., 2010. Hydraulic properties of fen peat soils in Poland. Geoderma 154(3-4), 188-195.

Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146(3-4), 403-411.

González Sánchez, A., Frausto Solís, J., Ojeda Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction.

Grassini, P., van Bussel, L.G., Van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., de Groot, H., van Ittersum, M.K., Cassman, K.G., 2015. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. Field Crops Research 177, 49-63.

Gray, J.M., Bishop, T.F., Yang, X., 2015. Pragmatic models for the prediction and digital mapping of soil properties in eastern Australia. Soil Research 53(1), 24-42.

Greene, R., Ford, G., 1985. The effect of gypsum on cation exchange in two red duplex soils. Soil Research 23(1), 61-74.

Gualtieri, J.A., Cromp, R.F., 1999. Support vector machines for hyperspectral remote sensing classification, 27th AIPR Workshop: Advances in Computer-Assisted Recognition. International Society for Optics and Photonics, pp. 221-233.

Gupta, S., Larson, W., 1982. Modeling Soil Mechanical behavior During Tillage 1. Predicting tillage effects on soil physical properties and processes (predictingtilla), 151-178.

Håkansson, I., 1990. A method for characterizing the state of compactness of the plough layer. Soil and tillage research 16(1-2), 105-120.

Håkansson, I., Reeder, R.C., 1994. Subsoil compaction by vehicles with high axle load—extent, persistence and crop response. Soil and Tillage Research 29(2), 277-304.

Hamza, M., Anderson, W., 2005. Soil compaction in cropping systems: A review of the nature, causes and possible solutions. Soil and tillage research 82(2), 121-145.

Han, H., Jiang, X., 2014. Overcome support vector machine diagnosis overfitting. Cancer informatics 13, CIN. S13875.

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

Hawkins, D.M., Basak, S.C., Mills, D., 2003. Assessing model fit by cross-validation. Journal of chemical information and computer sciences 43(2), 579-586.

Haykin, S., Network, N., 2004. A comprehensive foundation. Neural networks 2(2004), 41.

Hazelton, P., Murphy, B., 2016. Interpreting soil test results: What do all the numbers mean? CSIRO publishing.

Heath, R., 2018. Editorial to John Ralph Essay Competition 2018: Should society determine the right to farm? Farm Policy Journal 15(5), 2-3.

Hedley, C., Roudier, P., Yule, I., Ekanayake, J., Bradbury, S., 2013. Soil water status and water table depth modelling using electromagnetic surveys for precision irrigation scheduling. Geoderma 199, 22-29.

Hedley, C., Yule, I., 2009. A method for spatial prediction of daily soil water status for precise irrigation scheduling. Agricultural Water Management 96(12), 1737-1745.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124(3), 383-398.

Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120(1-2), 75-93.

Hengl, T., Rossiter, D.G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Soil Research 41(8), 1403-1422.

Henriksen, H.J., Barlebo, H.C., 2008. Reflections on the use of Bayesian belief networks for adaptive management. Journal of Environmental Management 88(4), 1025-1036.

Heuvelink, G.B., Brus, D.J., de Gruijter, J.J., 2006. Optimization of sample configurations for digital mapping of soil properties with universal kriging. Developments in soil science 31, 137-151.

Hiemstra, P., Hiemstra, M.P., 2013. Package 'automap'. compare 105, 10.

Hinneburg, A., Keim, D.A., 1998. An efficient approach to clustering in large multimedia databases with noise, KDD, pp. 58-65.

Hochman, Z., Gobbett, D., Holzworth, D., McClelland, T., van Rees, H., Marinoni, O., Garcia, J.N., Horan, H., 2013. Reprint of "Quantifying yield gaps in rainfed cropping systems: A case study of wheat in Australia". Field Crops Research 143, 65-75.

Holt, J., Mushobozi, W., Day, R., Knight, J., Kimani, M., Njuki, J., Musebe, R., 2006. A simple Bayesian network to interpret the accuracy of armyworm outbreak forecasts. Annals of Applied Biology 148(2), 141-146.

Holzworth, D.P., Huth, N.I., 2011. Simple software processes and tests improve the reliability and usefulness of a model. Environmental modelling & software 26(4), 510-516.

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM – Evolution towards a new generation of agricultural systems simulation. Environmental Modelling & Software 62, 327-350.

Hopmans, J., Minasny, B., Harter, T., 2003. Neural network prediction of soil hydraulic properties of alluvial soils, EGS-AGU-EUG Joint Assembly.

Horn, R., Fleige, H., 2003. A method for assessing the impact of load on mechanical stability and on physical properties of soils. Soil and Tillage Research 73(1-2), 89-99.

Houghes, G., 1968. On the mean accuracy of statistical pattern recognition. IEEE Trans. Inform. Theory 14(1), 55-63.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A practical guide to support vector classification.

Huang, J., McBratney, A.B., Malone, B.P., Field, D.J., 2018. Mapping the transition from pre-European settlement to contemporary soil conditions in the Lower Hunter Valley, Australia. Geoderma 329, 27-42.

Hudson, G., Wackernagel, H., 1994. Mapping temperature using kriging with external drift: theory and an example from Scotland. International journal of Climatology 14(1), 77-91.

Irmak, A., Jones, J., Batchelor, W., Irmak, S., Boote, K., Paz, J., 2006. Artificial neural network model as a data analysis tool in precision farming. Transactions of the ASABE 49(6), 2027-2037.

Isbell, R., NCST (2016)'The Australian soil classification.'. CSIRO Publishing: Melbourne.

Isbell, R., 1996. The Australian Soil Classification CSIRO Publishing. Collingwood, Australia.

Ishaq, M., Ibrahim, M., Lal, R., 2003. Persistence of subsoil compaction effects on soil properties and growth of wheat and cotton in Pakistan. Experimental Agriculture 39(4), 341-348.

Iyer, M.S., Rhinehart, R.R., 1999. A method to determine the required number of neural-network training repetitions. IEEE Transactions on Neural Networks 10(2), 427-432.

Jafarzadeh, A., Pal, M., Servati, M., FazeliFard, M., Ghorbani, M., 2016. Comparative analysis of support vector machine and artificial neural network models for soil cation exchange capacity prediction. International journal of environmental science and technology 13(1), 87-96.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters 31(8), 651-666.

Jayawardane, N., Blackwell, J., 1985. The effects of gypsum-enriched slots on moisture movement and aeration in an irrigated swelling clay. Soil Research 23(4), 481-492.

Jayawardane, N., Blackwell, J., Blackwell, P., 1988. Fragment size distribution within slots created in a duplex soil by a prototype rotary slotter. Soil and Tillage Research 12(1), 53-64.

Jenny, H., 1941. Factors ofSoil Formation, A System of Quantitative Pedology. McGraw-Hill.

John, R.S., Draper, N.R., 1975. D-optimality for regression designs: a review. Technometrics 17(1), 15-23.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. European journal of agronomy 18(3-4), 235-265.

Karimi, Y., Prasher, S., Madani, A., Kim, S., 2008. Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. Canadian Biosystems Engineering 50(7), 13-20.

Kassambara, A., 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning, 1. STHDA.

Kaundal, R., Kapoor, A.S., Raghava, G.P., 2006. Machine learning techniques in disease forecasting: a case study on rice blast prediction. BMC bioinformatics 7(1), 485.

Kazman, Z., Shainberg, I., Gal, M., 1983. Effect of low levels of exchangeable sodium and applied phosphogypsum on the infiltration rate of various soils 1. Soil Science 135(3), 184-192.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N., Meinke, H., Hochman, Z., 2003a. An overview of APSIM, a model designed for farming systems simulation. European journal of agronomy 18(3-4), 267-288.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N., Meinke, H., Hochman, Z., 2003b. An overview of APSIM, a model designed for farming systems simulation. European journal of agronomy 18(3), 267-288.

Keller, T., Arvidsson, J., 2004. Technical solutions to reduce the risk of subsoil compaction: effects of dual wheels, tandem wheels and tyre inflation pressure on stress propagation in soil. Soil and Tillage Research 79(2), 191-205.

Keller, T., Défossez, P., Weisskopf, P., Arvidsson, J., Richard, G., 2007. SoilFlex: A model for prediction of soil stresses and soil compaction due to agricultural field traffic including a synthesis of analytical approaches. Soil and Tillage Research 93(2), 391-411.

Kelly, N., Bennett, J.M., Starasts, A., 2017. Networked learning for agricultural extension: a framework for analysis and two cases. The Journal of Agricultural Education and Extension 23(5), 399-414.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11(1), 137-148.

Khaki, S., Wang, L., 2019. Crop Yield Prediction Using Deep Neural Networks. arXiv preprint arXiv:1902.02860.

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

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. European Journal of Soil Science 67(3), 276-284.

Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H., Shamshirband, S., Ab Hamid, S.H., 2015. Developing a fuzzy clustering model for better energy use in farm management systems. Renewable and Sustainable Energy Reviews 48, 27-34.

Kingwell, R., Fuchsbichler, A., 2011. The whole-farm benefits of controlled traffic farming: An Australian appraisal. Agricultural Systems 104(7), 513-521.

Kirby, J., 1991. Critical ‐ state soil mechanics parameters and their variation for Vertisols in eastern Australia. Journal of Soil Science 42(3), 487-499.

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.

Kleynhans, T., Montanaro, M., Gerace, A., Kanan, C., 2017. Predicting Top-of-Atmosphere Thermal Radiance Using MERRA-2 Atmospheric Data with Deep Learning. Remote Sensing 9(11), 1133.

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

Koekkoek, E., Booltink, H., 1999. Neural network models to predict soil water retention. European Journal of Soil Science 50(3), 489-495.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, Ijcai. Montreal, Canada, pp. 1137-1145.

Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering 160, 3-24.

Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154(3-4), 340-347.

Kristensen, K., Rasmussen, I.A., 2002. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. Computers and Electronics in Agriculture 33(3), 197-217.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, pp. 1097-1105.

Kubat, M., 2015. An Introduction to Machine Learning. Springer, Place of publication not identified.

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

Kuhn, M., Johnson, K., 2013. Applied predictive modeling, 26. Springer.

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

Kulkarni, S., Bajwa, S., Huitink, G., 2010. Investigation of the effects of soil compaction in cotton. Transactions of the ASABE 53(3), 667-674.

Kusumo, B., 2018. In Situ Measurement of Soil Carbon with Depth using Near Infrared (NIR) Spectroscopy, IOP Conference Series: Materials Science and Engineering. IOP Publishing, pp. 012235.

Kusumo, B.H., Hedley, C., Hedley, M., Hueni, A., Tuohy, M., Arnold, G., 2008. The use of diffuse reflectance spectroscopy for in situ carbon and nitrogen analysis of pastoral soils. Soil Research 46(7), 623-635.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296-311.

Lamorski, K., Pachepsky, Y., Sławiński, C., Walczak, R., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Science Society of America Journal 72(5), 1243-1247.

Lark, R., 2002. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma 105(1-2), 49-80.

Lark, R., Webster, R., 2006. Geostatistical mapping of geomorphic variables in the presence of trend. Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group 31(7), 862-874.

Larson, W., Gupta, S., Useche, R., 1980. Compression of Agricultural Soils from Eight Soil Orders 1. Soil Science Society of America Journal 44(3), 450-457.

Laslett, G., McBratney, A., Pahl, P.J., Hutchinson, M., 1987. Comparison of several spatial prediction methods for soil pH. Journal of Soil Science 38(2), 325-341.

Lawes, R., Oliver, Y., Robertson, M., 2009. Capturing the in-field spatial–temporal dynamic of yield variation. Crop and Pasture Science 60(9), 834-843.

LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., Sackinger, E., 1995. Comparison of learning algorithms for handwritten digit recognition, International conference on artificial neural networks. Perth, Australia, pp. 53-60.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water resources research 31(2), 387-398.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis & Prevention 40(4), 1611-1618.

Li, Y., 2010. Can the spatial prediction of soil organic matter contents at various sampling scales be improved by using regression kriging with auxiliary information? Geoderma 159(1-2), 63-75.

Li, Y., Shi, Z., Li, F., Li, H.-Y., 2007. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. Computers and Electronics in Agriculture 56(2), 174-186.

Liao, S.-H., Chu, P.-H., Hsiao, P.-Y., 2012. Data mining techniques and applications–A decade review from 2000 to 2011. Expert systems with applications 39(12), 11303-11311.

Lipiec, J., Hajnos, M., Świeboda, R., 2012. Estimating effects of compaction on pore size distribution of soil aggregates by mercury porosimeter. Geoderma 179, 20-27.

Lipiec, J., Medvedev, V., Birkas, M., Dumitru, E., Lyndina, T., Rousseva, S., Fulajtar, E., 2003. Effect of soil compaction on root growth and crop yield in Central and Eastern Europe. International agrophysics 17(2), 61-70.

Lisa, M., 2019. Lime and liming - managing soil health, GRDC Update Paper. The Grains Research and Development Corporation.

Littleboy, M., Silburn, D., Freebairn, D., Woodruff, D., Hammer, G., 1989. PERFECT-A computer simulation model of Productivity Erosion Runoff Functions to Evaluate Conservation Techniques. Bulletin-Queensland Department of Primary Industries (Australia).

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

Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment 164, 324-333.

Lobry de Bruyn, L., 2019. Learning opportunities: Understanding farmers' soil testing practice through workshop activities to improve extension support for soil health management. Soil Use and Management.

Lobry de Bruyn, L., Andrews, S., 2016. Are Australian and United States farmers using soil information for soil health management? Sustainability 8(4), 304.

Lobry de Bruyn, L., Jenkins, A., Samson-Liebig, S., 2017. Lessons learnt: sharing soil knowledge to improve land management and sustainable soil use. Soil Science Society of America Journal 81(3), 427-438.

Lobsey, C., Viscarra Rossel, R., 2016. Sensing of soil bulk density for more accurate carbon accounting. European Journal of Soil Science 67(4), 504-513.

Lobsey, C., Viscarra Rossel, R., Roudier, P., Hedley, C., 2017. rs ‐ local data ‐ mines information from spectral libraries to improve local calibrations. European journal of soil science 68(6), 840-852.

Lund, E., 1999. Veris Technologies. Saline, KS, 67401.

Lush, D., 2018. Should society determine the right to farm? Farm Policy Journal 15(4), 4.

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232, 34-44.

Malone, B.P., Odgers, N.P., Stockmann, U., Minasny, B., McBratney, A.B., 2018. Digital mapping of soil classes and continuous soil properties. Pedometrics, 373-413.

McBratney, A., de Gruijter, J., 1992a. A continuum approach to soil classification by modified fuzzy k‐means with extragrades. Journal of Soil Science 43(1), 159-175.

McBratney, A., de Gruijter, J., 1992b. A continuum approach to soil classification by modified fuzzy k‐means with extragrades. European Journal of Soil Science 43(1), 159-175.

McBratney, A., Pringle, M., 1999. Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. Precision Agriculture 1(2), 125-152.

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

McBratney, A., Webster, R., 1986. Choosing functions for semi‐variograms of soil properties and fitting them to sampling estimates. Journal of soil Science 37(4), 617-639.

McBratney, A., Webster, R., Burgess, T., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables—I: Theory and method. Computers & Geosciences 7(4), 331-334.

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

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109(1), 41-73.

McBratney, A.B., Odeh, I.O., 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77(2-4), 85-113.

McBratney, A.B., Odeh, I.O., Bishop, T.F., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97(3-4), 293-327.

Mcbratney, A.X., Whelan, B.M., Shatar, T.M., 2007. Variability and uncertainty in spatial, temporal and spatiotemporal crop‐yield and related data, Ciba Foundation Symposium 210‐Precision Agriculture: Spatial and Temporal Variability of Environmental Quality: Precision Agriculture: Spatial and Temporal Variability of Environmental Quality: Ciba Foundation Symposium 210. Wiley Online Library, pp. 141-160.

McDonald, R.C., Isbell, R., Speight, J.G., Walker, J., Hopkins, M., 1998. Australian soil and land survey: field handbook. CSIRO publishing.

McGarry, D., Sharp, G., Bray, S., 1999. The current status of soil degradation in Queensland cropping soils. Report No. DNRQ990092. Queensland Department of Natural Resources, Brisbane, Qld.

McHugh, A., Tullberg, J., Freebairn, D., 2009. Controlled traffic farming restores soil structure. Soil and Tillage Research 104(1), 164-172.

McKenzie, D., Abbott, T., Chan, K., Slavich, P., Hall, D., 1993. The nature, distribution and management of sodic soils in New-South-Wales. Soil Research 31(6), 839-868.

McKenzie, D., Bernardi, A., Chan, K., Nicol, H., Banks, L., Rose, K., 2002a. Sodicity v. yield decline functions for a Vertisol (Grey Vertosol) under border check and raised bed irrigation. Australian Journal of Experimental Agriculture 42(3), 363-368.

McKenzie, N., Bramley, R., Farmer, T., Janik, L., Murray, W., Smith, C., McLaughlin, M., 2003. Rapid soil measurement–a review of potential benefits and opportunities for the Australian grains industry. GRDC Project CSO 27.

McKenzie, N., Dixon, J., 2006. Monitoring Soil Condition Across Australia—Recommendations from the Expert Panels. Prepared on behalf of the National Coordinating Committee on Soil and Terrain, National Land and Water Resources Audit, Canberra.

McKenzie, N., Henderson, B., McDonald, W., 2002b. Monitoring Soil Change. Principles and Practices for Australian Conditions, CSIRO Land and Water, Technical Report 18/02, May 2002.

McKenzie, N.J., 1992. Soils of the Lower Macquarie Valley, New South Wales. CSIRO Division of Soils.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89(1-2), 67-94.

Mehta, P., Shah, H., Kori, V., Vikani, V., Shukla, S., Shenoy, M., 2015. Survey of unsupervised machine learning algorithms on precision agricultural data, Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on. IEEE, pp. 1-8.

Merdun, H., Çınar, Ö., Meral, R., Apan, M., 2006. Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. Soil and Tillage Research 90(1-2), 108-116.

Metternicht, G., Zinck, J., 2003. Remote sensing of soil salinity: potentials and constraints. Remote sensing of Environment 85(1), 1-20.

Milenova, B.L., Campos, M.M., 2002. O-cluster: Scalable clustering of large high dimensional data sets, 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE, pp. 290-297.

Minasny, B., Hopmans, J., Harter, T., Eching, S., Tuli, A., Denton, M., 2004. Neural networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data. Soil Science Society of America Journal 68(2), 417-429.

Minasny, B., McBratney, A., 2002. The neuro-m method for fitting neural network parametric pedotransfer functions. Soil Science Society of America Journal 66(2), 352-361.

Minasny, B., McBratney, A., 2006a. Latin hypercube sampling as a tool for digital soil mapping. Developments in soil science 31, 153-606.

Minasny, B., McBratney, A., 2010. Conditioned Latin hypercube sampling for calibrating soil sensor data to soil properties, Proximal soil sensing. Springer, pp. 111-119.

Minasny, B., McBratney, A.B., 2006b. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & geosciences 32(9), 1378-1388.

Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma 140(4), 324-336.

Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemometrics and Intelligent Laboratory Systems 94(1), 72-79.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301-311.

Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93(3-4), 225-253.

Mohanty, B.P., Cosh, M.H., Lakshmi, V., Montzka, C., 2017. Soil moisture remote sensing: State-of-the-science. Vadose Zone Journal 16(1).

Momma, M., Bennett, K.P., 2002. A pattern search method for model selection of support vector regression, Proceedings of the 2002 SIAM International Conference on Data Mining. SIAM, pp. 261-274.

Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., Mukhopadhyay, A., 2017. Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data. The Egyptian Journal of Remote Sensing and Space Science 20(1), 61-70.

Moran, M.S., Inoue, Y., Barnes, E., 1997. Opportunities and limitations for image-based remote sensing in precision crop management. Remote sensing of Environment 61(3), 319-346.

Mulla, D.J., 2013. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosystems engineering 114(4), 358-371.

Müller, W.G., 2001. Collecting Spatial Data: Optimum Design of Experiments for Random Fields, 2nd Ed.

Naidu, R., Merry, R.H., Churchman, G., Wright, M., Murray, R., Fitzpatrick, R.W., Zarcinas, B., 1993. Sodicity in South Australia-a review. Soil Research 31(6), 911-929.

Nari, K., Yang-Won, L., 2016. Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State. Journal of the Korean Society of Surveying Geodesy Photogrammetry and Cartography (34(4)), 383-390.

Nelson, M., Bishop, T., Triantafilis, J., Odeh, I., 2011. An error budget for different sources of error in digital soil mapping. European Journal of Soil Science 62(3), 417-430.

Neumann, K., Verburg, P.H., Stehfest, E., Müller, C., 2010. The yield gap of global grain production: A spatial analysis. Agricultural systems 103(5), 316-326.

Newlands, N.K., Townley-Smith, L., 2010. Predicting energy crop yield using bayesian networks, Proceedings of the Fifth IASTED International Conference, pp. 014-106.

Niang, M.A., Nolin, M.C., Jégo, G., Perron, I., 2014. Digital Mapping of soil texture using RADARSAT-2 polarimetric synthetic aperture radar data. Soil Science Society of America Journal 78(2), 673-684.

Niedbała, G., 2019. Simple model based on artificial neural network for early prediction and simulation winter rapeseed yield. Journal of integrative agriculture 18(1), 54-61.

Nuttall, J., Armstrong, R., 2010. Impact of subsoil physicochemical constraints on crops grown in the Wimmera and Mallee is reduced during dry seasonal conditions. Soil Research 48(2), 125-139.

Nuttall, J.G., Armstrong, R., Connor, D., 2003. Evaluating physicochemical constraints of Calcarosols on wheat yield in the Victorian southern Mallee. Australian Journal of Agricultural Research 54(5), 487-497.

O'sullivan, M., Henshall, J., Dickson, J., 1999. A simplified method for estimating soil compaction. Soil and Tillage Research 49(4), 325-335.

Odeh, I., Chittleborough, D., McBratney, A., 1992. Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Science Society of America Journal 56(2), 505-516.

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

Odeh, I.O., McBratney, A., Chittleborough, D., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67(3-4), 215-226.

Odeh, I.O., McBratney, A.B., 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. Geoderma 97(3-4), 237-254.

Odgers, N.P., McBratney, A.B., Minasny, B., 2015. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma 237, 190-198.

Oldfield, E.E., Bradford, M.A., Wood, S.A., 2019. Global meta-analysis of the relationship between soil organic matter and crop yields. Soil 5(1), 15-32.

Oliver, M., Webster, R., 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. Catena 113, 56-69.

Orton, T.G., Mallawaarachchi, T., Pringle, M.J., Menzies, N.W., Dalal, R.C., Kopittke, P.M., Searle, R., Hochman, Z., Dang, Y.P., 2018. Quantifying the economic impact of soil constraints on Australian agriculture: A case‐study of wheat. Land Degradation & Development 29(11), 3866-3875.

Oster, J., 1982. Gypsum usage in irrigated agriculture: a review. Fertilizer research 3(1), 73-89.

Oster, J., Jayawardane, N., 1998. Agricultural management of sodic soils.

Pachepsky, Y.A., Timlin, D., Varallyay, G., 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Science Society of America Journal 60(3), 727-733.

Panda, S.S., Ames, D.P., Panigrahi, S., 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sensing 2(3), 673-696.

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

Park, S., Hwang, C., Vlek, P., 2005. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. Agricultural Systems 85(1), 59-81.

Park, Y.R., Murray, T.J., Chen, C., 1996. Predicting sun spots using a layered perceptron neural network. IEEE Transactions on Neural Networks 7(2), 501-505.

Peng, J., Loew, A., Merlin, O., Verhoest, N.E., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. Reviews of Geophysics 55(2), 341-366.

Pennock, D., Yates, T., Braidek, J., 2007. Soil sampling designs. Soil sampling and methods of analysis, 1-14.

Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Pham, D.T., Dimov, S.S., Nguyen, C.D., 2005. Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 219(1), 103-119.

Pierce, F., Anderson, N., Colvin, T., Schueller, J., Humburg, D., McLaughlin, N., Sadler, E., 1997. Yield mapping. The state of site-specific management for agriculture.

Pollino, C.A., Henderson, C., 2010. Bayesian networks: A guide for their application in natural resource management and policy. Landscape Logic, Technical Report 14.

Prabhakar, M., Prasad, Y., Rao, M.N., 2012. Remote sensing of biotic stress in crop plants and its applications for pest management, Crop stress and its management: Perspectives and strategies. Springer, pp. 517-545.

Price, P., 2010. Preface: Combating Subsoil Constraints: R&D for the Australian grains industry. Soil Research 48(2), i-iii.

Pringle, M., McBratney, A., Whelan, B., Taylor, J., 2003. A preliminary approach to assessing the opportunity for site-specific crop management in a field, using yield monitor data. Agricultural Systems 76(1), 273-292.

Qian, N., 1999. On the momentum term in gradient descent learning algorithms. Neural networks 12(1), 145-151.

Qiao, D., Shi, H., Pang, H., Qi, X., Plauborg, F., 2010a. Estimating plant root water uptake using a neural network approach. Agricultural water management 98(2), 251-260.

Qiao, D.M., Shi, H.B., Pang, H.B., Qi, X.B., Plauborg, F., 2010b. Estimating plant root water uptake using a neural network approach. Agricultural Water Management 98(2), 251-260.

Queensland Government, 2018. SILO climate data.

Quinlan, J.R., 1992. Learning with continious classes, Australian Join Conference on Artifical Intelligence, Hobart, Vic, pp. 343-348.

Quirk, J., 2001. The significance of the threshold and turbidity concentrations in relation to sodicity and microstructure. Soil Research 39(6), 1185-1217.

Ramesh, V., Ramar, K., Babu, S., 2013. Parallel K-Means Algorithm on Agricultural Databases. IJCSI International Journal of Computer Science Issues 10(1), 1694-0814.

Rayment, G.E., Lyons, D.J., 2011. Soil chemical methods: Australasia, 3. CSIRO publishing.

Rengasamy, P., 2002. Transient salinity and subsoil constraints to dryland farming in Australian sodic soils: an overview. Australian Journal of Experimental Agriculture 42(3), 351-361.

Rengasamy, P., Greene, R., Ford, G., Mehanni, A., 1984. Identification of dispersive behaviour and the management of red-brown earths. Soil Research 22(4), 413-431.

Rengasamy, P., Olsson, K., 1991. Sodicity and soil structure. Soil Research 29(6), 935-952.

Roberton, S.D., Bennett, J.M., 2017. Efficacy of delaying cotton defoliation to mitigate compaction risk at wet harvest. Crop and Pasture Science 68(5), 466-473.

Robertson, D., Wang, Q.J., 2004. Bayesian networks for decision analyses — an application to irrigation system selection. Australian Journal of Experimental Agriculture 44(2), 145-150.

Robertson, M., Llewellyn, R., Mandel, R., Lawes, R., Bramley, R., Swift, L., Metz, N., O'Callaghan, C., 2012. Adoption of variable rate fertiliser application in the Australian grains industry: status, issues and prospects. Precision Agriculture 13(2), 181-199.

Robinson, N., Rampant, P., Callinan, A., Rab, M., Fisher, P., 2009. Advances in precision agriculture in south-eastern Australia. II. Spatio-temporal prediction of crop yield using terrain derivatives and proximally sensed data. Crop and Pasture Science 60(9), 859-869.

Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. Science 344(6191), 1492-1496.

Rossel, R.V., Adamchuk, V., Sudduth, K., McKenzie, N., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time, Advances in agronomy. Elsevier, pp. 243-291.

Roudier, P., Beaudette, D., Hewitt, A., 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. Digital soil assessments and beyond, 227-231.

Roudier, P., Hedley, C., Lobsey, C., Rossel, R.V., Leroux, C., 2017. Evaluation of two methods to eliminate the effect of water from soil vis–NIR spectra for predictions of organic carbon. Geoderma 296, 98-107.

Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., Plümer, L., 2010. Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance. Computers and Electronics in Agriculture 74(1), 91-99.

Ruß, G., 2009. Data mining of agricultural yield data: A comparison of regression models, Industrial Conference on Data Mining. Springer, pp. 24-37.

Ruß, G., Kruse, R., 2011. Exploratory hierarchical clustering for management zone delineation in precision agriculture, Industrial Conference on Data Mining. Springer, pp. 161-173.

Sacks, W.J., Deryng, D., Foley, J.A., Ramankutty, N., 2010. Crop planting dates: an analysis of global patterns. Global Ecology and Biogeography 19(5), 607-620.

Sankaran, S., Mishra, A., Ehsani, R., Davis, C., 2010. A review of advanced techniques for detecting plant diseases. Computers and Electronics in Agriculture 72(1), 1-13.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T., 2016. Meta-learning with memory-augmented neural networks, International conference on machine learning, pp. 1842-1850.

Sarani, F., Ahangar, A.G., Shabani, A., 2016. Predicting ESP and SAR by artificial neural network and regression models using soil pH and EC data (Miankangi Region, Sistan and Baluchestan Province, Iran). Archives of Agronomy and Soil Science 62(1), 127-138.

Sattler, C., Nagel, U.J., Werner, A., Zander, P., 2010. Integrated assessment of agricultural production practices to enhance sustainable development in agricultural landscapes. Ecological Indicators 10(1), 49-61.

Schaap, M.G., Bouten, W., 1996. Modeling water retention curves of sandy soils using neural networks. Water Resources Research 32(10), 3033-3040.

Schaap, M.G., Leij, F.J., Van Genuchten, M.T., 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Science Society of America Journal 62(4), 847-855.

Schierhorn, F., Faramarzi, M., Prishchepov, A.V., Koch, F.J., Müller, D., 2014. Quantifying yield gaps in wheat production in Russia. Environmental Research Letters 9(8), 084017.

Schoier, G., Borruso, G., 2004. A clustering method for large spatial databases, International Conference on Computational Science and Its Applications. Springer, pp. 1089-1095.

Sedaghatpour, S., Ellis, T., Hignett, C., Bellotti, B., 1995. Six years of controlled traffic cropping research on a red brown earth at Roseworthy in South Australia, Proceedings 1st National Controlled Traffic Conference, pp. 13-14.

Seif, G., 2018. The 5 Clustering Algorithms Data Scientists Need to Know.

Shainberg, I., Keren, R., Frenkel, H., 1982. Response of Sodic Soils to Gypsum and Calcium Chloride Application 1. Soil Science Society of America Journal 46(1), 113-117.

Shainberg, I., Rhoades, J., Prather, R., 1981. Effect of Low Electrolyte Concentration on Clay Dispersion and Hydraulic Conductivity of a Sodic Soil 1. Soil Science Society of America Journal 45(2), 273-277.

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

Shaw, R., Brebber, L., Ahern, C., Weinand, M., 1994. A review of sodicity and sodic soil behavior in Queensland. Soil Research 32(2), 143-172.

Sibson, R., 1981. A brief description of natural neighbour interpolation. Interpreting multivariate data.

Simbahan, G.C., Dobermann, A., 2006. Sampling optimization based on secondary information and its utilization in soil carbon mapping. Geoderma 133(3-4), 345-362.

Sirjacobs, D., Hanquet, B., Lebeau, F., Destain, M.-F., 2002. On-line soil mechanical resistance mapping and correlation with soil physical properties for precision agriculture. Soil and Tillage Research 64(3-4), 231-242.

Smith, C., Johnston, M., Lorentz, S., 1997. Assessing the compaction susceptibility of South African forestry soils. II. Soil properties affecting compactibility and compressibility. Soil and Tillage Research 43(3-4), 335-354.

Smith, C.S., Howes, A.L., Price, B., McAlpine, C.A., 2007a. Using a Bayesian belief network to predict suitable habitat of an endangered mammal–The Julia Creek dunnart (Sminthopsis douglasi). Biological Conservation 139(3-4), 333-347.

Smith, C.S., Howes, A.L., Price, B., McAlpine, C.A., 2007b. Using a Bayesian belief network to predict suitable habitat of an endangered mammal–The Julia Creek dunnart (Sminthopsis douglasi). Biological Conservation 139(3), 333-347.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and computing 14(3), 199-222.

Stafford, J., Ambler, B., Lark, R., Catt, J., 1996. Mapping and interpreting the yield variation in cereal crops. Computers and Electronics in Agriculture 14(2-3), 101-119.

Steeneveld, W., van der Gaag, L.C., Barkema, H.W., Hogeveen, H., 2010. Simplify the interpretation of alert lists for clinical mastitis in automatic milking systems. Computers and electronics in agriculture 71(1), 50-56.

Stephanie, 2016. What is Hierarchical Clustering?

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

Suykens, J.A., Vandewalle, J., 2000. Recurrent least squares support vector machines. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 47(7), 1109-1114.

Tamari, S., Wösten, J., Ruiz-Suarez, J., 1996. Testing an artificial neural network for predicting soil hydraulic conductivity. Soil Science Society of America Journal 60(6), 1732-1741.

Tari, F., 1996. A Bayesian Network for predicting yield response of winter wheat to fungicide programmes. Computers and Electronics in Agriculture 15(2), 111-121.

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

Taylor, J.H., 1983. Benefits of permanent traffic lanes in a controlled traffic crop production system. Soil and Tillage Research 3(4), 385-395.

Tennant, D., Scholz, G., Dixon, J., Purdie, B., 1992. Physical and chemical characteristics of duplex soils and their distribution in the south-west of Western Australia. Australian Journal of Experimental Agriculture 32(7), 827-843.

Thomasson, J.A., Baillie, C.P., Antille, D.L., Lobsey, C.R., McCarthy, C.L., 2019. Autonomous Technologies in Agricultural Equipment: A Review of the State of the Art. American Society of Agricultural and Biological Engineers.

Tilling, A.K., O'Leary, G.J., Ferwerda, J.G., Jones, S.D., Fitzgerald, G.J., Rodriguez, D., Belford, R., 2007. Remote sensing of nitrogen and water stress in wheat. Field Crops Research 104(1-3), 77-85.

Tranter, G., Minasny, B., McBratney, A., 2010. Estimating Pedotransfer Function Prediction Limits Using Fuzzy k-Means with Extragrades All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. Soil Science Society of America Journal 74(6), 1967-1975.

Troldborg, M., Aalders, I., Towers, W., Hallett, P.D., McKenzie, B.M., Bengough, A.G., Lilly, A., Ball, B.C., Hough, R.L., 2013. Application of Bayesian Belief Networks to quantify and map areas at risk to soil threats: Using soil compaction as an example. Soil and Tillage Research 132, 56-68.

Tsiropoulos, Z., Fountas, S., Gemtos, T., Gravalos, I., Paraforos, D., 2013. Management information system for spatial analysis of tractor-implement draft forces, Precision agriculture'13. Springer, pp. 349-356.

Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology 49(11), 1225-1231.

Tullberg, J., Antille, D.L., Bluett, C., Eberhard, J., Scheer, C., 2018. Controlled traffic farming effects on soil emissions of nitrous oxide and methane. Soil and Tillage Research 176, 18-25.

Tullberg, J., Yule, D., McGarry, D., 2007. Controlled traffic farming—from research to adoption in Australia. Soil and Tillage Research 97(2), 272-281.

Twarakavi, N.K., Šimůnek, J., Schaap, M., 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. Soil Science Society of America Journal 73(5), 1443-1452.

Üstün, B., Melssen, W.J., Oudenhuijzen, M., Buydens, L.M.C., 2005. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. Analytica Chimica Acta 544(1), 292-305.

Van Bergeijk, J., Goense, D., 1996. Soil tillage resistance as tool to map soil type differences. Precision Agriculture (precisionagricu3), 605-616.

Van Der Gaag, L.C., Bolt, J., Loeffen, W., Elbers, A., 2010. Modelling patterns of evidence in Bayesian networks: a case-study in classical swine fever, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, pp. 675-684.

Van Egmond, F., Loonstra, E., Limburg, J., 2010. Gamma ray sensor for topsoil mapping: The Mole, Proximal Soil Sensing. Springer, pp. 323-332.

Vapnik, V., 1995. The nature of statistical learning theory. J. Wiley & Sons, New York.

Varis, O., 1997. Bayesian decision analysis for environmental and resource management. Environmental Modelling & Software 12(2), 177-185.

Vašát, R., Heuvelink, G., Borůvka, L., 2010. Sampling design optimization for multivariate soil mapping. Geoderma 155(3-4), 147-153.

Vellido, A., Martín-Guerrero, J.D., Lisboa, P.J., 2012. Making machine learning models interpretable, ESANN. Citeseer, pp. 163-172.

Viscarra Rossel, R.A., Adamchuk, V., Sudduth, K., McKenzie, N., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time, Advances in agronomy. Elsevier, pp. 243-291.

Viscarra Rossel, R.A., Bouma, J., 2016. Soil sensing: A new paradigm for agriculture. Agricultural Systems 148, 71-74.

Viscarra Rossel, R.A., Lobsey, C.R., Sharman, C., Flick, P., McLachlan, G., 2017. Novel Proximal Sensing for Monitoring Soil Organic C Stocks and Condition. Environmental Science & Technology 51(10), 5630-5641.

Viscarra Rossel, R.A., McBratney, A., 1998. Laboratory evaluation of a proximal sensing technique for simultaneous measurement of soil clay and water content. Geoderma 85(1), 19-39.

Viscarra Rossel, R.A., McBratney, A., Minasny, B., 2010a. Proximal soil sensing.

Viscarra Rossel, R.A., McBratney, A.B., Minasny, B., 2010b. Proximal soil sensing. Springer Science & Business Media.

Voorhees, W., Young, R., Lyles, L., 1979. Wheel traffic considerations in erosion research. Transactions of the ASAE 22(4), 786-0790.

Waiser, T.H., Morgan, C.L., Brown, D.J., Hallmark, C.T., 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Science Society of America Journal 71(2), 389-396.

Walvoort, D.J., Brus, D., De Gruijter, J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Computers & Geosciences 36(10), 1261-1267.

Wani, M.A., Bhat, F.A., Afzal, S., Khan, A.I., 2019. Advances in Deep Learning. Springer.

Warrick, A.W., 2001. Soil physics companion. CRC press.

Webster, R., Lark, M., 2012. Field sampling for environmental science and management. Routledge.

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

Wendelberger, J.G., 1981. The Computation of Laplacian Smoothing Splines with Examples, WISCONSIN UNIV-MADISON DEPT OF STATISTICS.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52, 394-403.

Wessels, L.F., Barnard, E., 1992. Avoiding false local minima by proper initialization of connections. IEEE Transactions on Neural Networks 3(6), 899-905.

Whelan, B., McBratney, A., 2000. The "null hypothesis" of precision agriculture management. Precision Agriculture 2(3), 265-279.

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

Whish, J., Butler, G., Castor, M., Cawthray, S., Broad, I., Carberry, P., Hammer, G., McLean, G., Routley, R., Yeates, S., 2005. Modelling the effects of row configuration on sorghum yield reliability in north-eastern Australia. Australian Journal of Agricultural Research 56(1), 11-23.

White, H., 1989. Learning in artificial neural networks: A statistical perspective. Neural computation 1(4), 425-464.

White, P.J., Broadley, M.R., 2003. Calcium in plants. Annals of botany 92(4), 487-511.

Widrow, B., Lehr, M.A., 1990. 30 years of adaptive neural networks: perceptron, Madaline, and backpropagation. Proceedings of the IEEE 78(9), 1415-1442.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wong, M., Wittwer, K., Oliver, Y., Robertson, M., 2010. Use of EM38 and gamma ray spectrometry as complementary sensors for high-resolution soil property mapping, Proximal soil sensing. Springer, pp. 343-349.

Wösten, J., Finke, P., Jansen, M., 1995. Comparison of class and continuous pedotransfer functions to generate soil hydraulic characteristics. Geoderma 66(3-4), 227-237.

Xu, X., Ester, M., Kriegel, H.-P., Sander, J., 1998. A distribution-based clustering algorithm for mining in large spatial databases, Proceedings 14th International Conference on Data Engineering. IEEE, pp. 324-331.

Yaduvanshi, N., Sharma, D., 2008. Tillage and residual organic manures/chemical amendment effects on soil organic matter and yield of wheat under sodic water irrigation. Soil and tillage research 98(1), 11-16.

Yan, L., Zhou, S., Feng, L., 2007. Delineation of Site-Specific Management Zones Based on Temporal and Spatial Variability of Soil Electrical Conductivity1. Pedosphere 17(2), 156-164.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A., Hann, S., Burt, J.E., Qi, F., 2011. Updating conventional soil maps through digital soil mapping. Soil Science Society of America Journal 75(3), 1044-1053.

Young, R., Wilson, B., Harden, S., Bernardi, A., 2009. Accumulation of soil carbon under zero tillage cropping and perennial vegetation on the Liverpool Plains, eastern Australia. Soil Research 47(3), 273-285.

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

Zhang, Y., Biswas, A., Ji, W., Adamchuk, V.I., 2017. Depth-specific prediction of soil properties in situ using vis-NIR spectroscopy. Soil Science Society of America Journal 81(5), 993-1004.

Zhao, G., Bryan, B.A., King, D., Luo, Z., Wang, E., Bende-Michl, U., Song, X., Yu, Q., 2013. Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing. Environmental Modelling & Software 41, 231-238.

Zhou, A., Zhou, S., Cao, J., Fan, Y., Hu, Y., 2000. Approaches for scaling DBSCAN algorithm to large spatial databases. Journal of computer science and technology 15(6), 509-526.

Zhu, A., Liu, J., Du, F., Zhang, S., Qin, C., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. European Journal of Soil Science 66(3), 535-547.

Zribi, M., Baghdadi, N., Nolin, M., 2011. Remote sensing of soil. Applied and Environmental Soil Science 2011.

9. Appendix

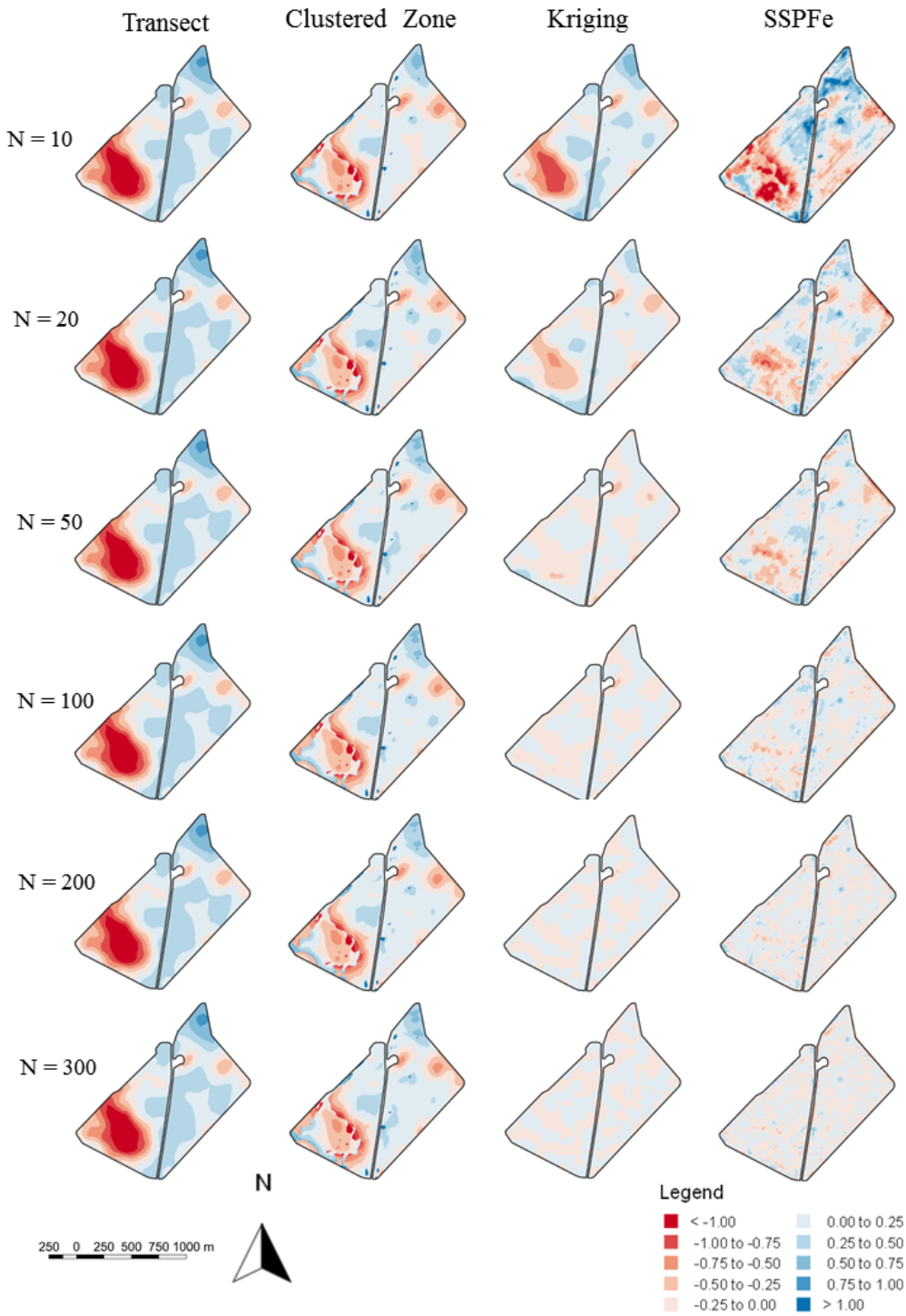


Figure 9.1 Prediction error maps of the 4 methods investigated for pH at 0 – 10 cm depth increment. Error maps shown for sampling densities N = 10, 20, 50, 100, 200 and 300. Red shades represent under prediction whilst blue shades represent over prediction.

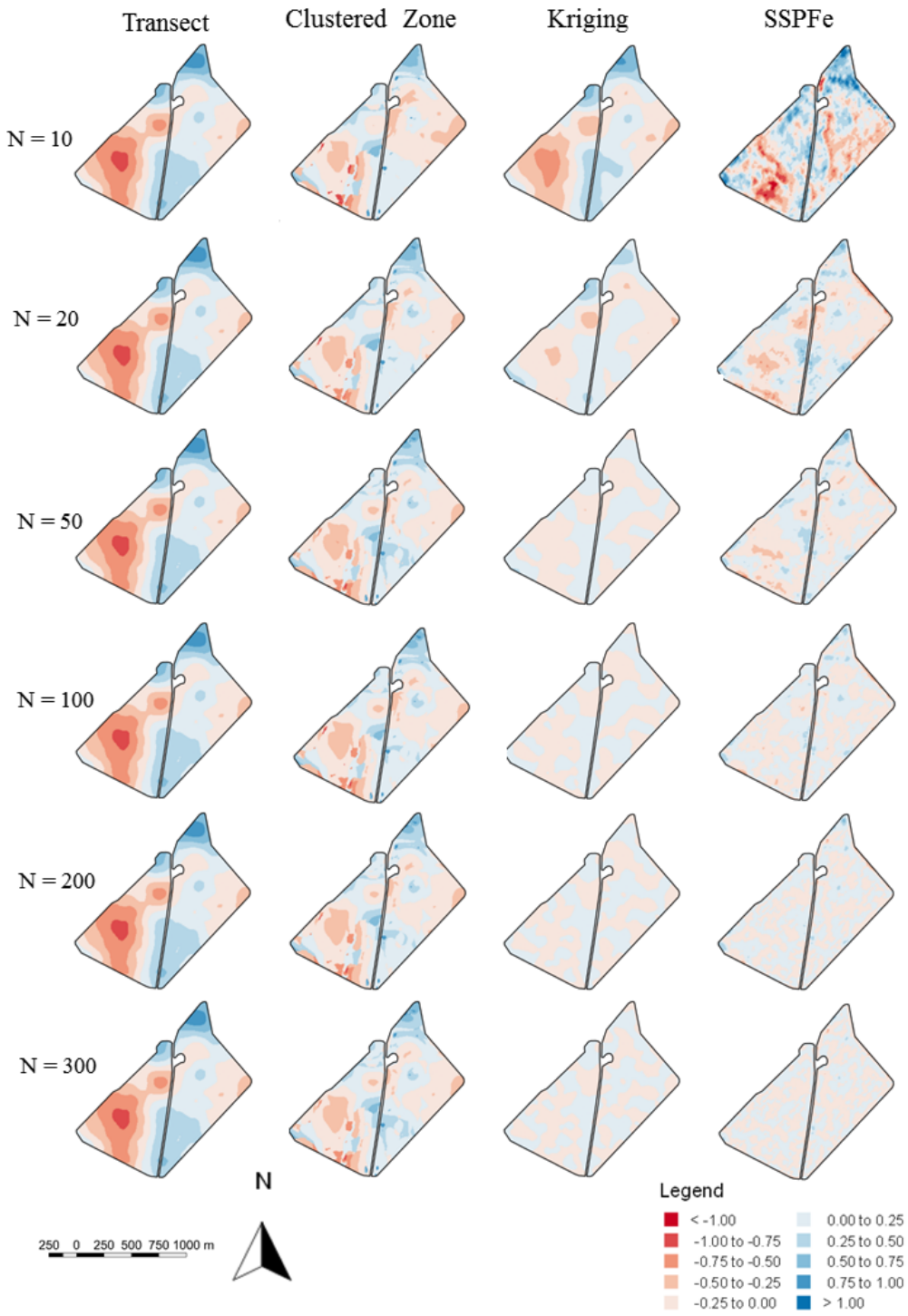


Figure 9.2 Prediction error maps of the 4 methods investigated for pH at 10 – 20 cm depth increment. Error maps shown for sampling densities N = 10, 20, 50, 100, 200 and 300. Red shades represent under prediction whilst blue shades represent over prediction.

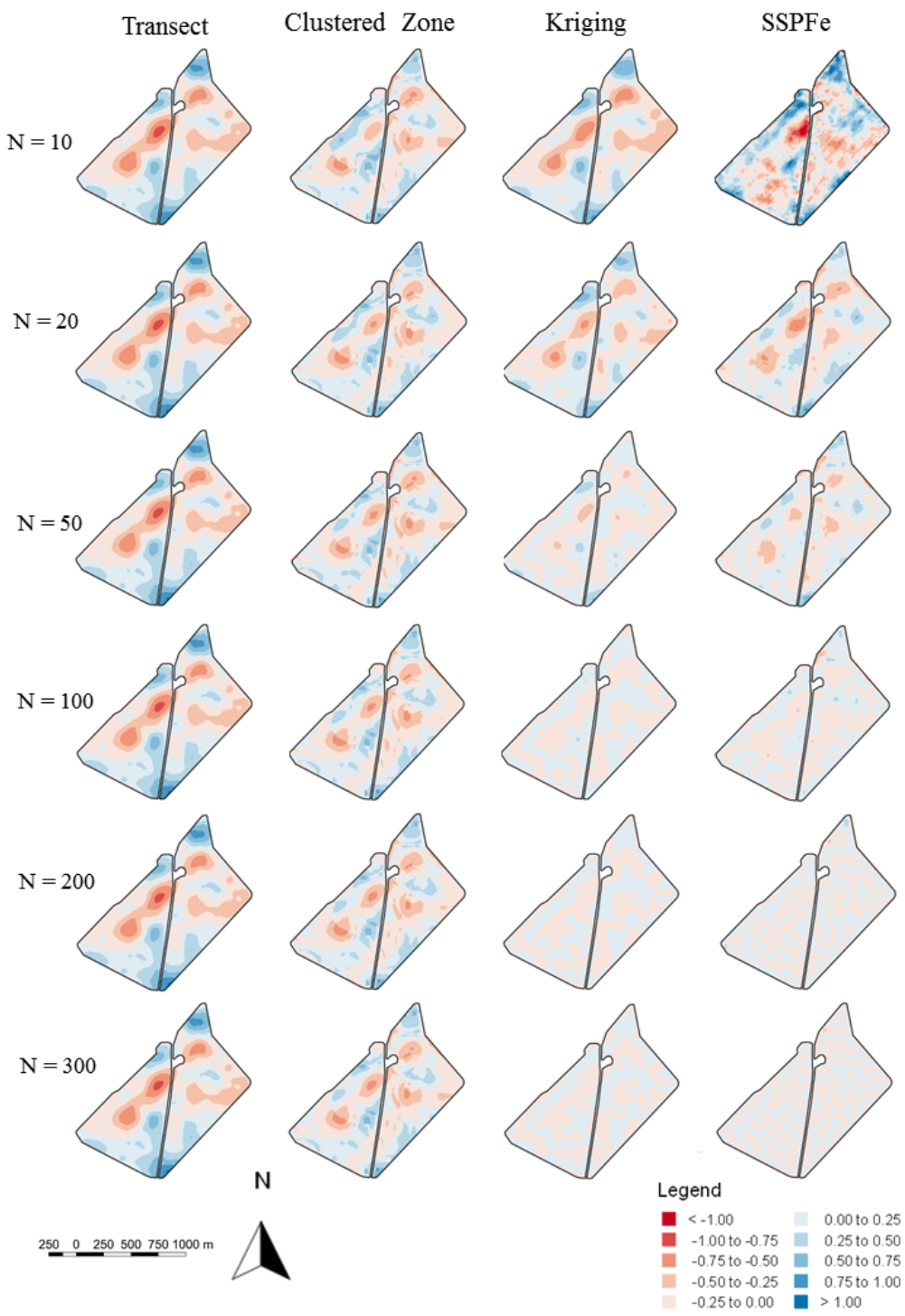


Figure 9.3 Prediction error maps of the 4 methods investigated for pH at 20 – 40 cm depth increment. Error maps shown for sampling densities N = 10, 20, 50, 100, 200 and 300. Red shades represent under prediction whilst blue shades represent over prediction.

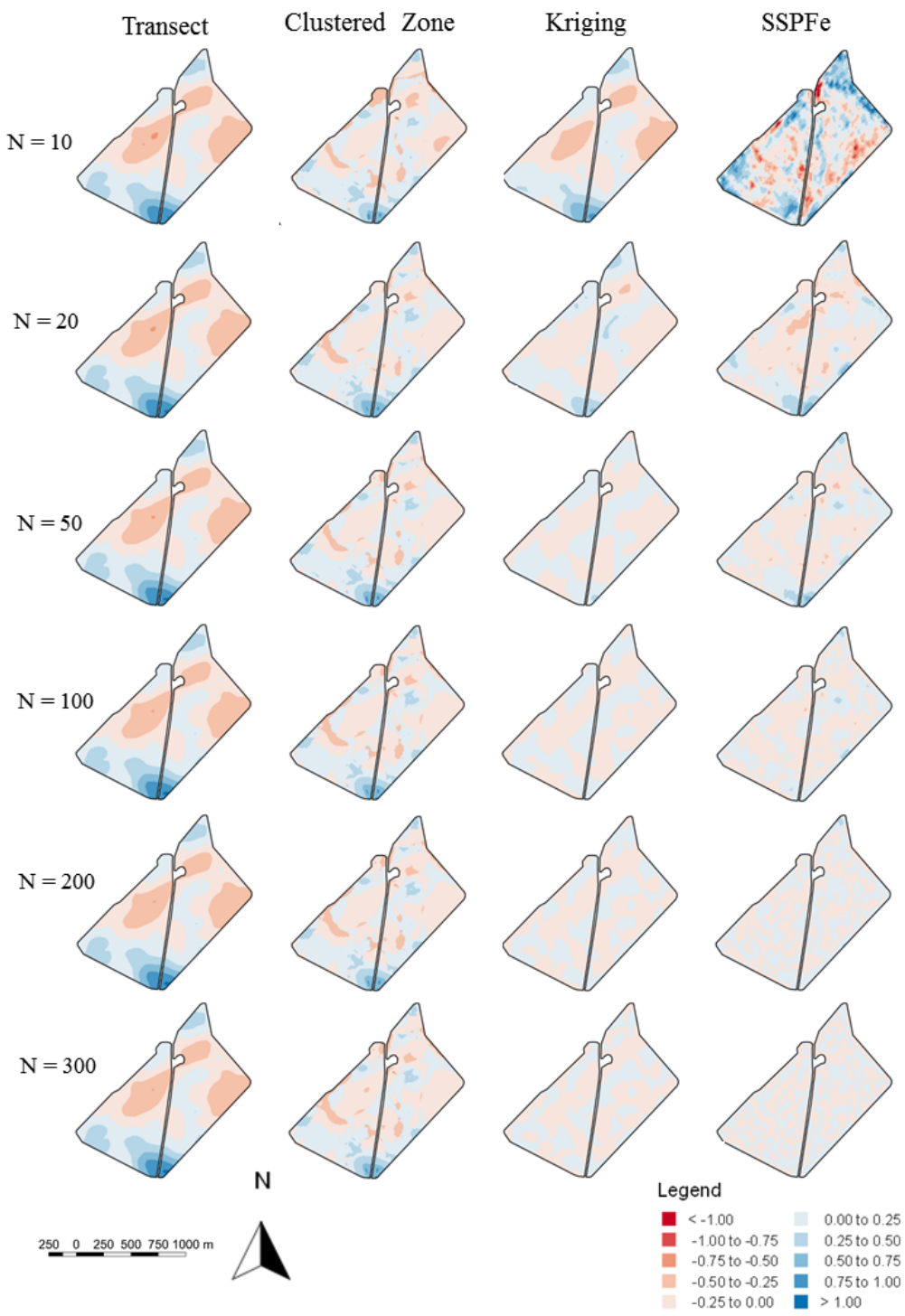


Figure 9.4 Prediction error maps of the 4 methods investigated for pH at 40 - 60 cm depth increment. Error maps shown for sampling densities N = 10, 20, 50, 100, 200 and 300. Red shades represent under prediction whilst blue shades represent over prediction.

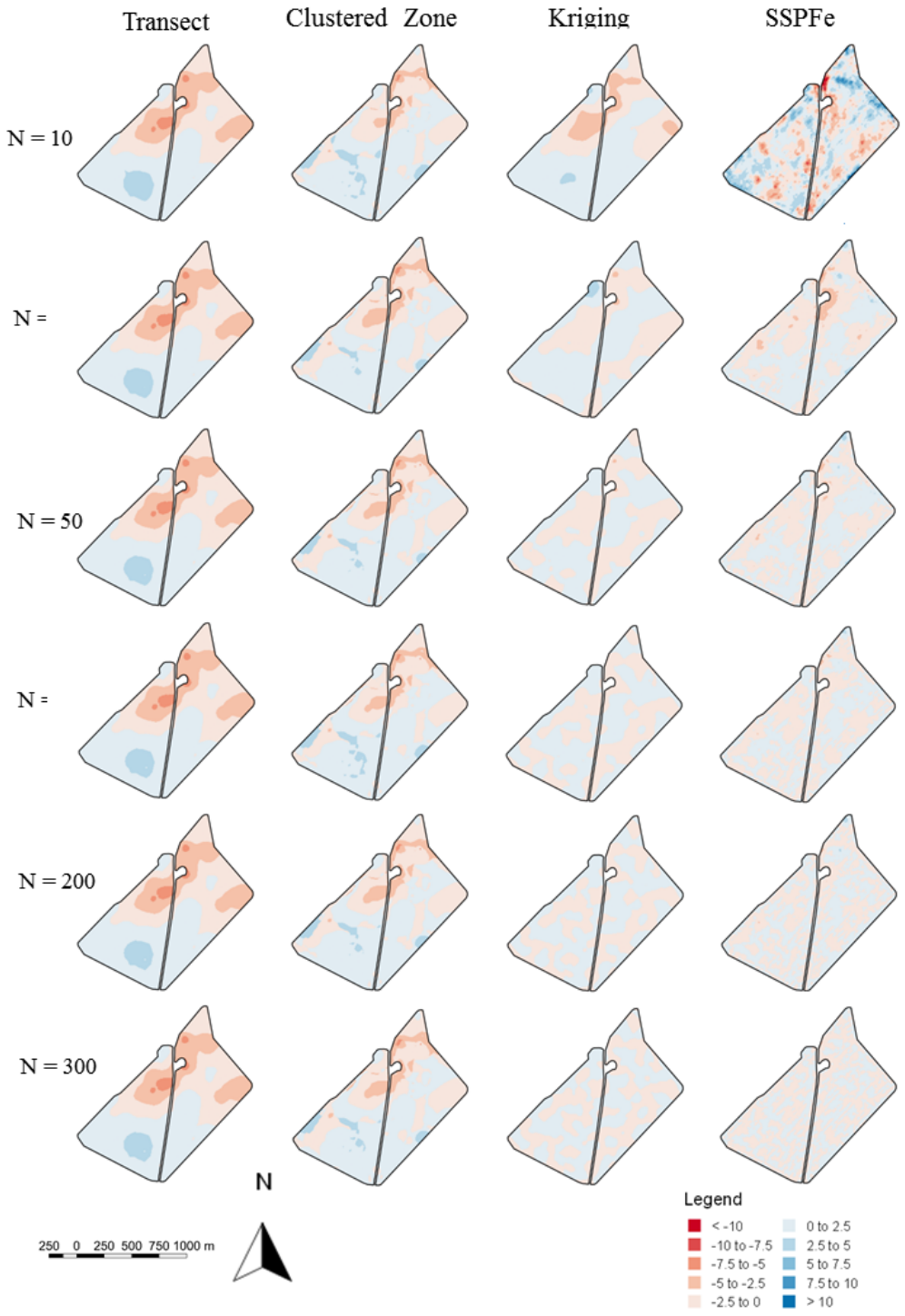


Figure 9.5 Prediction error maps of the 4 methods investigated for ESP at 0 – 10 cm depth increment. Error maps shown for sampling densities N = 10, 20, 50, 100, 200 and 300. Red shades represent under prediction whilst blue shades represent over prediction.

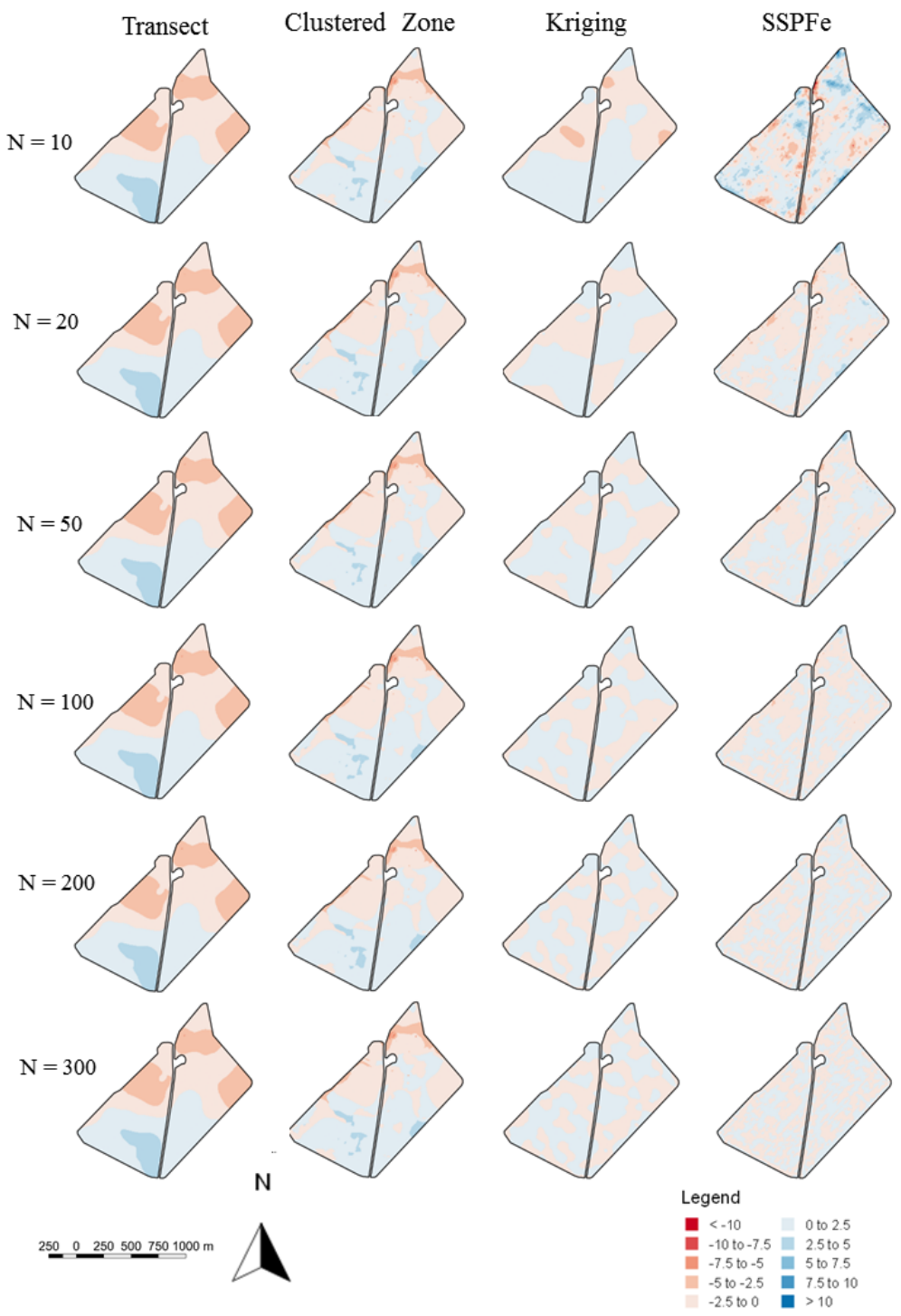


Figure 9.6 Prediction error maps of the 4 methods investigated for ESP at 10 - 20 cm depth increment. Error maps shown for sampling densities N = 10, 20, 50, 100, 200 and 300. Red shades represent under prediction whilst blue shades represent over prediction.

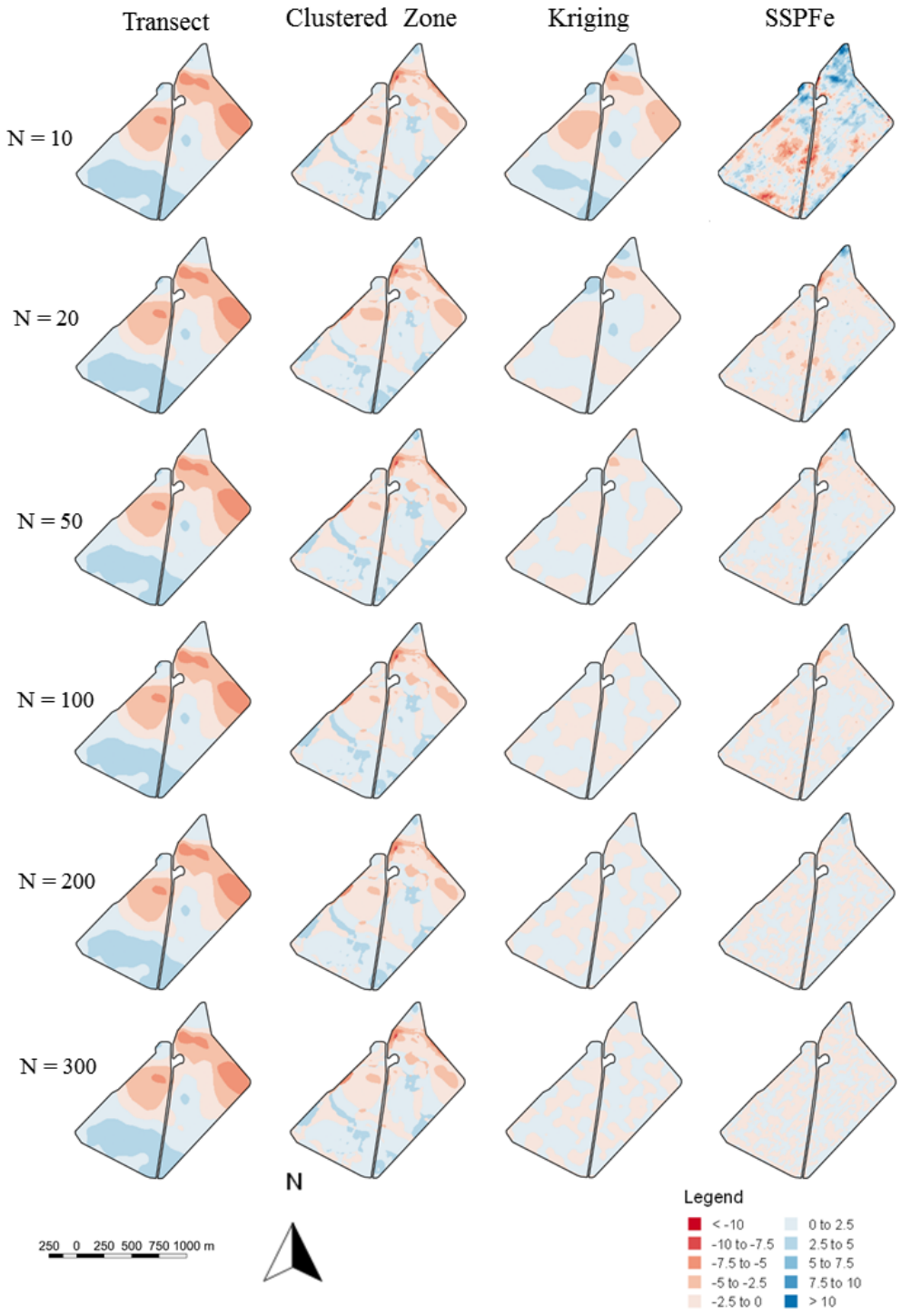


Figure 9.7 Prediction error maps of the 4 methods investigated for ESP at 20 – 40 cm depth increment. Error maps shown for sampling densities N = 10, 20, 50, 100, 200 and 300. Red shades represent under prediction whilst blue shades represent over prediction.

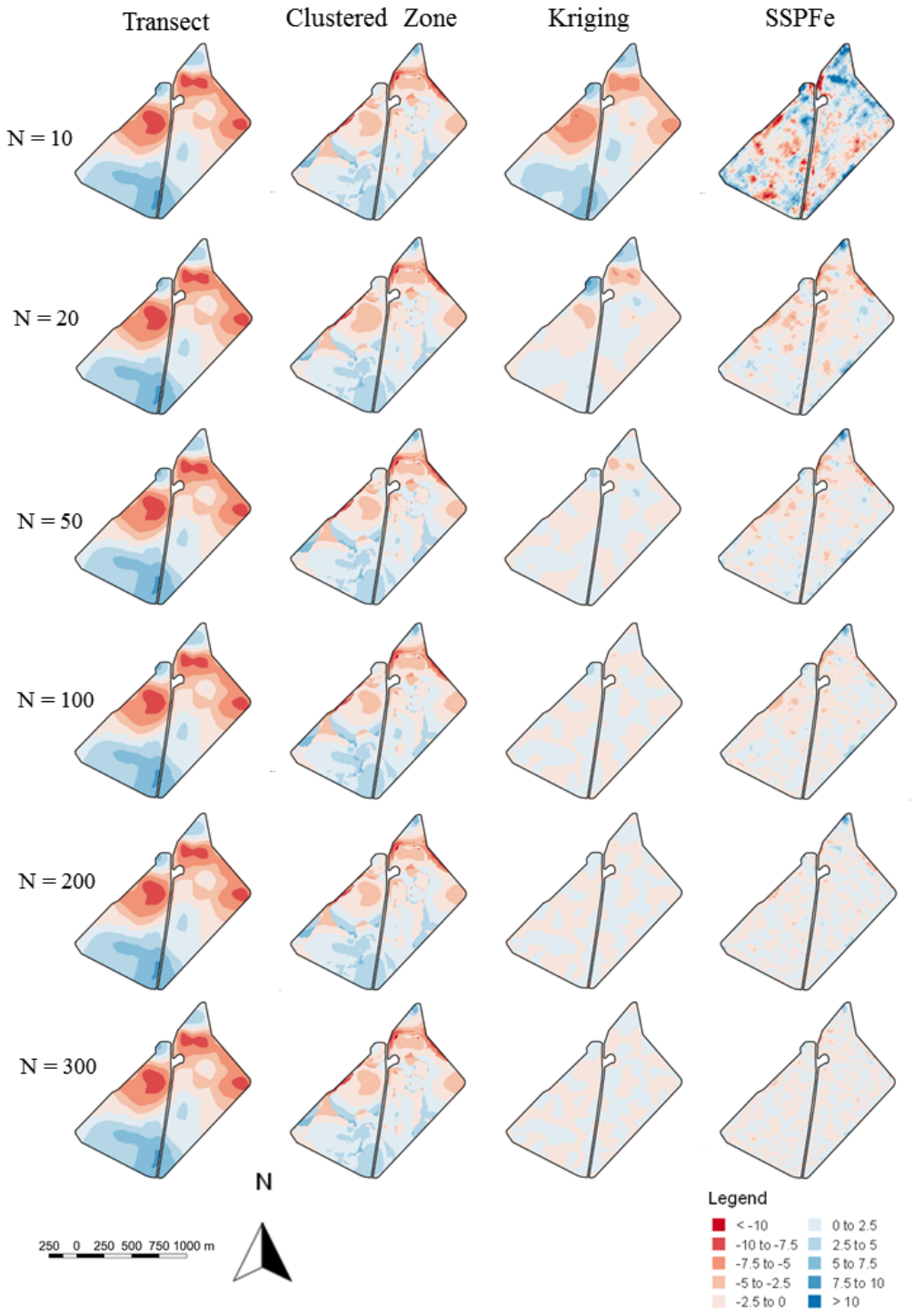


Figure 9.8 Prediction error maps of the 4 methods investigated for ESP at 40 - 60 cm depth increment. Error maps shown for sampling densities N = 10, 20, 50, 100, 200 and 300. Red shades represent under prediction whilst blue shades represent over prediction