Optimization of Sampling Designs for Validating Digital Soil Maps

Yakun Zhang

Department of Natural Resource Sciences
McGill University, Montreal
April 2016

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Master of Science

ABSTRACT

Meeting food demand for ever increasing global population can be attained through sustainable management of soil resources. This requires a thorough understanding of soil properties and processes and calls for methods to quantify and display spatial variability of soil. Three dimensional digital soil mapping (3D-DSM) with its ability to quantify both the horizontal and the vertical variability has become popular in recent days. The state-of-the-art data mining techniques including 3D regression kriging (RK) has been used to uncover complex soil-landscape relationships but not assessed at small scales. In addition, recent advances in proximal soil sensing allow measurement and prediction of various soil properties simultaneously and rapidly at multiple depths and provide required information for DSM. Furthermore, sampling design (SD) plays a vital role in providing a reliable input for DSM, whereas its effectiveness on 3D-DSM has not been tested.

A total of 148 sample locations, identified by six SDs, including grid sampling (GS), grid random sampling (GRS), simple random sampling (SRS), stratified random sampling (StRS), transect sampling (TS), and conditioned Latin hypercube sampling (cLHS), were used to collect vis-NIR spectra data to about 1-m depth *in-situ* using a commercial soil profiler from a small agricultural farm in Macdonald campus, McGill University. A subset of 32 sample locations were identified to collect soil cores down to 1-m depth and sampled at 10-cm depth intervals. A total of 251 samples were analyzed in laboratory for a range of soil properties. Partial least square regression was used to develop soil-spectral relationship model. Predicted soil and uncertainty maps for soil properties were developed using 3D-DSM with RK from the calibration dataset (103 locations) and assessed using validation dataset (45 locations). Further three regression techniques, including generalized linear model (GLM), regression tree (RT), and random forest (RF) were tested and compared for accuracy and efficiency. Maps developed using sub samples (45 locations) identified by six SDs were further compared with the original map produced by the full dataset (148 locations) and individually validated by the rest 103 locations.

The results showed that a good prediction was obtained for soil organic matter (SOM) and water-related soil properties from *in-situ* vis-NIR spectra, while a fair prediction was obtained for other properties. RF outperformed GLM and RT by quantifying the non-linear soil-landscape relationship, displaying weak spatial structure of regression residuals, and resulting in a more robust prediction model with high accuracy and low uncertainty. The predicted maps clearly

presented the soil spatial variability, reflected the interactions among soil properties, and displayed the associated soil forming processes. Among the SDs, StRS with both good spatial and feature space coverage better represented the distribution of original maps and showed a small prediction uncertainty, while cLHS produced higher validation accuracy. SRS resulted in good validation results, while requires further exploration for its robustness. The main contribution of this thesis was to assess and optimize the methods and techniques for 3D-DSM and associated SDs and quantify both the horizontal and vertical variability of multiple soil properties.

RÉSUMÉ

La réponse à la demande alimentaire pour une population mondiale croissante peut être atteinte à travers une gestion durable des ressources du sol. Ceci exigerait une compréhension des propriétés et des processus du sol et nécessiterait des méthodes de quantification de la variabilité du sol. La cartographie numérique à trois dimensions (CNS-3D) a une capacité de quantifier à la fois la variabilité horizontale et verticale s'est répandue dernièrement. Les dernières techniques incluant la régression à trois dimensions dite 'Kriging' (RK) a été utilisée pour explorer la relation complexe sol-paysage et non pour une évaluation à petite échelle. En outre, les récents progrès dans la détection proximale du sol permettent de mesurer et de prédire simultanément et rapidement les différentes propriétés du sol à multiples profondeurs et fournissent les informations nécessaires pour la cartographie numérique du sol (CNS). De plus, le plan d'échantillonnage (PE) joue un rôle essentiel en fournissant un apport solide pour la cartographie numérique à trois dimensions, alors que son efficacité à trois dimensions n'a pas été testée.

Sur le terrain agricole du campus Mcdonald à l'université McGill, un total de 148 sites d'échantillonnage, identifiés par six SDs incluant une grille d'échantillonnage (GE), une grille d'échantillonnage aléatoire (GEA), un échantillonnage aléatoire simple (EAS), un échantillonnage aléatoire stratifié (EASt), un échantillonnage transect (ET) et échantillonnage Hypercube latin (EHL), ont été utilisés *in situ* pour recueillir des données du spectre vis-NIR à environ 1 m de profondeur avec un profileur de sol. Un sous-ensemble de 32 sites d'échantillonnage a été identifié pour un carottage de sol jusqu'à 1 m de profondeur avec un intervalle de 10 cm. Un total de 251 échantillons a été analysé au laboratoire pour une gamme de propriétés de sol. Un modèle de régression partielle des moindres carrés a été utilisé pour développer un modèle de la relation sol-spectre. La prédiction de la carte et des propriétés du sol a été développée en utilisant la cartographie numérique à trois dimensions associée à la régression 'kriging' de l'ensemble des données d'étalonnage (103 sites) et évalués par la validation de l'ensemble des données (45 sites). Pour la précision et l'efficacité, trois autres techniques incluant le modèle linéaire généralisé (GLM), l'arbre à régression (AR) et la forêt aléatoire (FA) ont été testées et comparées.

Les résultats ont montré qu'une bonne prédication a été obtenue pour la matière organique du sol et l'eau en relation avec les propriétés du sol à partir du spectre vis-NIR *in situ*, tandis qu'une prédiction juste a été obtenue pour les autres propriétés. La forêt aléatoire a dépassé le modèle linéaire généralisé et l'arbre à régression en quantifiant la relation non linéaire sol-paysage,

affichant une faible structure spatiale de la régression résiduelle, et résultant en un modèle de prédiction plus robuste avec une grande précision et une faible incertitude. Les cartes prédites présentent clairement la variabilité spatiale du sol, reflétant les interactions entre les propriétés du sol, et ont affiché les processus associés à la formation du sol. Parmi les différents PE, les EASt ayant à la fois une bonne couverture spatiale et une caractéristique dans l'espace représentent mieux la distribution des cartes originales et montrent une petite incertitude de prédiction, alors que l'EHL a démontré une grande précision de validation. L'EAS a aboutit à de bons résultats de validation, alors qu'il nécessite une exploration plus poussée pour sa robustesse. La contribution principale de cette thèse était d'évaluer et d'optimiser les méthodes et techniques pour la cartographie numérique à trois dimensions et les plans d'échantillonnage associés, et de quantifier la variabilité horizontale et verticale les différentes propriétés du sol.

ACKNOWLEDGEMENTS

This project was funded by Natural Science and Engineering Research Council (NSERC) of Canada. Student's scholarship and travel grants from McGill University are also highly acknowledged.

I would like to express sincere thanks to Duminda, Savitoz, and Bharath who helped me with field sample collection in the cold winter 2014, and to Hélène Lalande who gave me high-quality guidance during lab measurement for almost 20 soil properties and 300 samples, and to the students in the Precision Agriculture group who collected and compiled the environmental data, and to Wenjun for her help with processing the spectral data, and to Hsin-Hui Huang who generously helped me with every problem I came across in GIS, and also to my advisory committee- Prof. Adamchuk for his insightful comments and encouragement. Without their support and collaboration, it would become very difficult to finish such a big project.

Furthermore, I am so grateful for meeting so many close friends in Montreal. I am so grateful for the constant thoughtfulness and support both in study and life from Ting Liu, Amy Li, Hua Cong, and Yue Wang. I thank Ebrahim, Nina, and Mi for accompanying me home at night after a whole day's work in last semester. I would like to thank Fei Tang who prepared delicious and nutritional meals many times for me in last month when I was busy with the thesis writing. I also thank Zhor Abail and Aboulkacem Lemtiri for translating the thesis abstract. Their help and support over the last two years have been invaluable.

In addition, a very special thanks to Dr. Tomislav Hengl and his colleagues, even though they do not know me. The YouTube digital soil mapping courses and GSIF tutorials did help me a lot in interpreting the statistical theory and producing the maps. Thanks for sharing all the useful knowledge, courses, and programming code to the public and for the effort they did to promote the development and advances in this field.

Importantly, I would like to thank my dear supervisor Prof. Asim Biswas. I am so impressed for his enthusiasm and dedication to his work and his students, and this greatly influenced me on my attitude to work. He is so strict and picky on his work and also my work, and he wants everything to be perfect. This inspired me to do better and better in my work. Besides, I am also grateful for his patience and tolerance on me for the last two years. I am very happy and honored to be his student, and it is really comfortable, interesting, and helpful to work with him.

Finally, I would like to thank my parents who love me and support me unconditionally all the time.

PREFACE AND CONTRIBUTIONS OF AUTHORS

This thesis was organized in a manuscript-based structure. This followed the guidelines set by Graduate and Postdoctoral Studies Office, McGill University. In Chapter 1, a brief introduction was given for the soil function, importance of understanding spatial variability of soil properties for sustainable soil management, development of 3D digital soil mapping and its challenges and opportunities for further work, and sampling design associated studies. The research objectives were proposed in this chapter. In Chapter 2, a comprehensive review of different sampling designs, including rationale, strength and weakness, application, and comparison, as well as a small review of digital soil mapping techniques was reported. All the research results were sub-divided and presented into 4 sections, including Chapter 3, Chapter 4, Chapter 5, and Appendix A, with prefaces before each chapter to show the connections between chapters and contributions of co-authors. These sections have been formatted into several research papers, which are either submitted or about to be submitted for publication in peer reviewed international journals. In Chapter 3, *in-situ* visible near infrared spectroscopy was tested for its ability to predict various soil properties at multiple depths and the developed spectral models were used to exhaustively obtain soil information for DSM. In Chapter 4, three regression techniques for regression kriging were assessed and compared, and the 3D-DSM products were discussed and presented for multiple soil properties at a field-scale. In Chapter 5, sub maps developed from a small sample size identified by different sampling designs were displayed and compared in order to choose the optimized and efficient sampling design. In Appendix A, a sigmoid model (profile depth function) was proposed to quantify the vertical distribution of soil pH in the local dataset, and the model was assessed for its generality by a global dataset with 432 soil profiles. In Chapter 6, an overall conclusion was reported and several future directions were identified.

The author of this thesis is the first author for all the manuscripts and takes charge of all the design and execution of the original research, including data collection and analysis, model development and assessment, interpretation of results, and manuscripts preparation. Prof. Asim Biswas is the thesis supervisor. He was highly involved in the whole study and was responsible for providing scientific comments and suggestions and technical assistance throughout every stage of the two-year study, reviewing and editing all the manuscripts and final thesis. And Prof. Biswas is the co-author and corresponding author for all the manuscripts. Prof. Viacheslav I. Adamchuk is the thesis advisory committee. He was in charge of the environmental covariates collection and compilation,

provided technical facilities for hyperspectral data and soil sample collection in this thesis, and continuously tracked the progress and gave suggestions for data analysis. Prof. Adamchuk is the co-author for four manuscripts due to his constructive suggestions to these manuscripts and his knowledge and technical support. Hélène Lalande is the laboratory technician. She gave much guidance and assistance during lab measurements for almost 20 soil properties and 300 soil samples. Duminda, Savitoz, and Bharath are graduate students in Prof. Biswas's lab and they helped with the soil samples and spectral data collection. Wenjun is a postdoc fellow in Prof. Adamchuk's lab and she supported the research with raw spectral data processing and advised on the spectral model development. Wenjun is co-author of three manuscripts due to her processing and advice on spectral data analysis.

The manuscript-based chapters are presented in the following order:

Chapter 2. Yakun Zhang & Asim Biswas. A review of sampling designs for calibrating digital soil maps. To be submitted to *Catena* (Impact factor: 2.61).

Chapter 3. Yakun Zhang, Asim Biswas, Wenjun Ji, & Viacheslav I. Adamchuk. Depth specific prediction of soil properties in-situ using vis-NIR spectroscopy. To be submitted to *Soil Science Society of America Journal* (Impact factor: 1.75).

Chapter 4. Yakun Zhang, Asim Biswas, Wenjun Ji, & Viacheslav I. Adamchuk. Three-dimensional digital soil mapping of multiple soil properties at a field-scale using 3D regression kriging. To be submitted to *Geoderma* (Impact factor: 2.85).

Chapter 5. Yakun Zhang, Asim Biswas, Wenjun Ji, & Viacheslav I. Adamchuk. Comparison of sampling designs for calibrating three-dimensional digital soil maps. To be submitted to *Geoderma* (Impact factor: 2.85).

Appendix A. Yakun Zhang, Asim Biswas, & Viacheslav I. Adamchuk. A sigmoid depth function to describe variations in soil pH in agricultural fields. Revised and resubmitted to *Geoderma* on Jun. 17th 2016 (Impact factor: 2.85).

CONTRIBUTION TO KNOWLEDGE

This thesis highlighted the following scientific contributions that mainly assessed and improved the current technology and methods to quantify both the horizontal and the vertical variability of soil properties:

1) A new profile depth function was proposed to quantify the vertical distribution of soil pH based on the understanding of the pedological and management features of agricultural fields. In addition, the generality of this model was tested for a global dataset. This is the first time that the sigmoid model has been developed for quantifying soil vertical variability and could be further used in 3D-DSM of soil pH.
2) The feasibility of *in-situ* vis-NIR spectra to predict a set of soil physical and chemical properties was tested down to 1m depth. The results showed a good prediction for SOM and water-related soil properties and fair prediction for various soil cations. This is noteworthy because none of previous studies have measured so many soil properties and reached 1m depth. In addition, this study enriched the assessment and application of *in-situ* vis-NIR spectra.
3) 3D regression kriging, an emerging method to produce 3D maps, has not been widely used for multiple soil properties at a small-scale. Therefore, we adopted it in this study and assessed the effectiveness of 3D RK in 3D-DSM for multiple soil properties. Additionally, three regression techniques were tested and compared to select the most proper one (random forest) for 3D RK. This provided valuable insights on the further selection and application of techniques for 3D-DSM.
4) Six different sampling designs optimized either in geographical space or feature space were compared and identified the optimized sampling designs for 3D regression kriging method. None of the previous studies have discussed the contribution and influence of different sampling designs on the accuracy of 3D maps in multiple soil layers. The results in this thesis suggested that a small sample selected by stratified random sampling was more efficient to represent the original distribution, and conditioned Latin hypercube sampling was highly recommended due to its high flexibility of optimization criteria and high accuracy of final maps.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

3D	Three-dimensional	DSM	Digital soil mapping
SD	Sampling design	GS	Grid sampling
SRS	Simple random sampling	SyRS	Systematic random sampling
StRS	Stratified random sampling	SCS	Spatial coverage sampling
GRS	Grid random sampling	TS	Transect sampling
CS	Cluster sampling	NS	Nested sampling
LHS	Latin hypercube sampling	cLHS	Conditioned LHS
wecLHS	Weighted LHS	DLHS	LHS with D-optimality
SSAS	Spatial simulated annealing sampling	SA	Simulated annealing
FKMS	Fuzzy k-means sampling	FCMS	Fuzzy c-means sampling
FKME	FKMS with extragrades	StSS	Stratified spatial sampling
RSS	Response surface sampling	KSS	Kennard-Stone sampling
MaxKV	Maximum kriging variance	AKV	Average kriging variance
PCA	Principal component analysis	iPCA	Iterative PCA
Vis-NIR	Visible near infrared	MIR	Mid-infrared
EC	Electrical conductivity	PC	Principal component
ECa	Apparent Electrical Conductivity	CTI	Compound topographic index
ECe	EC of the saturation extract	VWC	Volumetric water content
NDVI	Normalized difference vegetation index	GWC	Gravimetric water content
EAQSF	Equal-area quadratic spline function	BD	Bulk density
MLR	Multiple linear regression	SOM	Soil organic matter
PLSR	Partial least square regression	SOC	Soil organic carbon
PCR	Principal component regression	LOI	Loss on ignition
GLM	Generalized linear model	P	Phosphorus
LOOCV	Leave-one-out cross validation	K	Potassium
RMSE	Root mean squared error	Na	Sodium
AIC	Akaike Information Criterion	R^2	Coefficient of determination

Mg	Magnesium	Zn	Zinc
CSR	Complete spatial randomness	Mn	Manganese
GP	Genetic programming	RT	Regression tree
RBF	Radial basis function	RF	Random forest
RK	Regression kriging	NF	Number of factors
ME	Mean error	RPD	Relative percent deviation
GSIF	Global soil information facilities	GPS	Global position system
RTK	Real time kinematic	2D	Two-dimensional
GIS	Geographic information system	OK	Ordinary kriging
U	Uranium	UK	Universal kriging
Th	Thorium	A	Absorbance
Cs	Caesium	R	Reflectance
TC	Total count	Sd	Standard deviation
CEC	Cation exchange capacity	CV	Cross-validation
SNV	Standard normal variate	SG	Savitzky-Golay filter
Al	Aluminum	Fe	Iron
EDF	Exponential decay function	TR	Tikhonov regularization
AWC	Available water capacity	Ca	Calcium
PS	Purposive sampling	CEC	Cation exchange capacity
MSSD	Mean squared shortest distance	CR	Continuum-removal
EMI	Electromagnetic induction	DEM	Digital elevation modal

CHAPTER 1
INTRODUCTION

Soil is an unconsolidated layer on the earth surface that supports all terrestrial life. It is an integral component of the global ecosystem and closely interacts with atmosphere, biosphere, hydrosphere, and lithosphere by transferring matter and energy. Soil plays a vital role in providing food and materials, maintaining biological diversity, activity, and productivity, regulating water and nutrients dynamics, storing carbon, filtering and buffering, supporting the civil structure, and preserving the cultural heritage (Lal and Shukla, 2004).

Initially weathered from rocks, soil has undergone long-term and complex pedogenic processes, leading to a diverse and distinct morphology. Various environmental factors, including climate, organism, relief, parent material, and time (Jenny, 1941), promoted the formation and intensified the distinction, resulting in high spatial variability of soil properties. The information on the spatial variability of soil properties greatly assists agricultural management, environmental policy making, and natural resource management. In addition, increasing global issues and challenges including population growth and heavy demand on food supplies, accelerated environmental degradation and soil erosion, and depletion in non-sustainable natural resources, have already more or less influenced soil to carry out its function and will exacerbate the impact with time. This calls for more sustainable management of soil resources which requires a thorough understanding of soil properties and processes and their spatial and temporal variation.

Traditional soil maps rely heavily on broad soil measurement and subjective judgement of surveyors with qualitative soil maps as final products which are usually inefficient, inaccurate, and lack predictive ability. Digital soil mapping (DSM) bears closely on the soil-landscape relationship and produces more detailed soil maps from exhaustive environmental variables by rigorous statistical methods. DSM substantially increases the efficiency of the mapping procedure and allows more accurate and quantitative prediction of soil properties at any location. Furthermore, with the discovery of soil anisotropy feature and advancements of computational methods, three-dimensional digital soil mapping (3D-DSM) has been inquired to display both the horizontal and the vertical variability of soil properties. It is not an easy task to thoroughly disclose the heterogeneous relationship and quantify the variation.

The current 3D mapping procedures implement either a combination of a profile depth function with a 2D interpolation technique or 3D geostatistical methods. The first method separates the

vertical and horizontal variation and could not allow a full representation of 3D relationship (Liu et al., 2016). In addition, the commonly used equal-area quadratic spline function increases the limitation on understanding the soil physical condition and other specific depth function e.g. exponential decay function restricts it on specific soil properties and lacks generality. 3D geostatistical methods with great mathematical advances have been adopted for global soil grid maps (Hengl et al., 2014). The proposed regression-kriging is an effective mapping method that simultaneously conducts the regression between soil properties and environmental covariates and interpolation of regression residuals. However, it has not been widely applied. This method also provides an opportunity to utilize multiple regression techniques but has been rarely tested. This inspires the work in this study to conduct a 3D-DSM by regression kriging for multiple soil properties at a field-scale and assess the multiple linear and non-linear regression techniques to discover the soil-landscape relationship.

Proximal soil sensing rapidly and accurately predicts soil properties at multiple depths and works as an alternative to traditional laborious laboratory measurement of soil properties. It has been used in DSM to assist in the soil data collection and simplify the DSM work (Brodsky et al., 2013). Various soil properties such as soil organic matter and clay content are fundamental constitutes of soil and showed strong and clear absorption features in vis-NIR band and could be predicted from vis-NIR spectra with high accuracy (Rossel and Lark, 2009). In addition, soil water was also well predicted due to the recognizable feature (Stenberg et al., 2010). However, other soil properties with either positive or negative relationship were observed with various prediction results and lacked robustness. In addition, due to the technical constraints and complex and uneasily controlled conditions, vis-NIR spectra has not been widely used *in-situ*. Therefore, an opportunity exists to assess the effectiveness and stability of the prediction of multiple soil properties from *in-situ* measurement of vis-NIR spectra. And this would further contribute to the 3D-DSM study.

Sampling design systematically selects a small set of samples according to specific criteria to provide an effective and reliable input for DSM. Different sampling designs have been adopted in DSM such as grid sampling, stratified random sampling, transect sampling, and conditioned Latin hypercube sampling. In addition, comparison and improvement has been made and continues to optimize the sampling designs for DSM. However, with the development of 3D-DSM, few studies have explored the feasibility and effectiveness of these sampling designs on capturing the variability in multiple soil layers and the accuracy of the 3D maps.

Therefore, the objectives of this thesis were: 1) assessing the feasibility and accuracy of *in-situ* vis-NIR spectra to predict multiple soil physical and chemical properties; 2) comparing and identifying the most effective regression technique for 3D regression kriging to quantify the complex soil-landscape relationship; 3) developing and testing the 3D-DSM with regression kriging method for multiple soil properties at a field-scale and interpreting the horizontal and vertical variability of the soil properties with soil forming processes and field condition; 4) optimizing the sampling designs for calibrating the 3D-DSM.

A set of studies were designed and conducted in an agricultural field in McGill University and the results were illustrated in the following chapters of this thesis. This thesis was formatted in a manuscript-based structure. In Chapter 2, a comprehensive review of different sampling designs, including rationale, strength and weakness, application, and comparison was reported and followed by a short review of digital soil mapping technique, especially 3D-DSM. All the original research results were organized and presented in Chapter 3, Chapter 4, Chapter 5, and Appendix A, with prefaces before each chapter to show the connections between chapters and contributions of co-authors. In Chapter 3, the feasibility and accuracy of using vis-NIR spectra to predict soil properties was demonstrated. In Chapter 4, three regression techniques for regression kriging were assessed and compared, and finally, the 3D-DSM products were discussed and presented for multiple soil properties at a field-scale. In Chapter 5, sub maps developed from a small sample size identified by six different sampling designs were displayed and compared in order to choose the optimized and efficient sampling design. In Appendix A, a sigmoid model (profile depth function) was proposed to quantify the vertical distribution of soil pH in the local dataset, and the model was assessed for its generality by a global dataset with 432 soil profiles. These chapters have been formatted into several research papers, which are either submitted or to be submitted for publication in peer reviewed international journals. In Chapter 6, an overall conclusion was reported and several future directions were identified.

PREFACE TO CHAPTER 2

Chapter 1 provided an overall picture of soil function and the necessity of a comprehensive understanding of the soil spatial variability for agricultural and environmental management and policy-making in the global context. In addition, it discussed the development from traditional soil maps and digital soil mapping to the three-dimensional digital soil mapping, their features and limitations, and an important step of DSM- sampling designs. In the end, it proposed three objectives of the whole thesis. The proposed objectives were carefully designed and performed in a set of studies, and the results of these studies would be presented in Chapter 3, Chapter 4, Chapter 5, and Appendix A. In order to conduct the experimental and statistical design in this study, a literature review about sampling designs and mapping techniques were made and presented in Chapter 2. First of all, a comprehensive review of sampling designs were given, including the rationale, advantages, disadvantages, and application of those sampling designs as well as future works suggested by this review. This helped to identify the sampling designs in Chapter 5 and compare the results obtained in Chapter 5 with previous literatures. Furthermore, a short review of different mapping techniques, especially 3D-DSM techniques, were demonstrated in Chapter 2, this assisted in selecting the appropriate methods in Chapter 4.

Chapter 2 has been written as a review paper format and will be submitted to *Catena* (impact factor: 2.61). The detailed information for authors is shown below:

Order of authors: Yakun Zhang[a] & Asim Biswas[a,*].

The author of the thesis solely reviewed and wrote the manuscript. Prof. Biswas provided advice on the review and edited the manuscript.

In addition, Dr. Richard Webster gave much scientific and beneficial advice on this review.

CHAPTER 2

A review of sampling designs for calibrating digital soil maps

Abstract:

Sampling design plays a crucial role in providing a reliable input for DSM and increasing the DSM efficiency. Sampling design with a predetermined sample size by considering budget and spatial variability, is a selection procedure for a set of sample locations either by spreading sample locations in geographical space or obtaining a good feature space coverage. A good feature space coverage ensures an accurate estimation of regression parameters, while a spatial coverage contributes to an effective spatial interpolation. This study firstly reviewed several statistical sampling designs and geometric sampling designs which mainly optimize the sampling pattern in geographical space and illustrated the strength and weakness of these sampling designs by considering the spatial coverage, simplicity, accuracy, and efficiency. Furthermore, Latin hypercube sampling which obtains a full representation of multivariate distribution was demonstrated in detail for its development, improvement, and application. In addition, fuzzy k-means sampling, response surface sampling, and Kennard-Stone sampling which optimize sampling pattern in feature space were also presented in this review. We then discussed the practical application issues which were mainly addressed by conditioned Latin hypercube sampling with the flexibility and feasibility of adding multiple optimization criteria. Moreover, as an important stage of the DSM, different validation methods were discussed and suggested that an independent dataset selected from probability sampling was superior for its free model assumption. For future work, we recommended: 1) exploring the sampling designs with both good spatial coverage and feature space coverage; 2) uncovering the real impacts of a sampling design on the integral DSM procedure; and 3) testing the feasibility and contribution of sampling designs in 3D-DSM with multiple layers variability.

2.1 Introduction

Soil survey, incorporating field sampling, laboratory analysis, data processing, and mapping, aims at classifying soil types and soil attributes of a specific field, with soil maps as ultimate products (McBratney et al., 2000b). In conventional soil surveys, soil maps were produced by qualitative delineation of soil boundaries based on understanding of soil forming factors (Jenny, 1941), which were greatly affected by subjective judgement and practical experience of surveyors (Clifford et

al., 2014). Therefore, the information conveyed by traditional soil maps is usually qualitative and relatively subjective and fails to allow a good predictive quality.

Increasing demand on soil information to solve a variety of agricultural issues, including site-specific management of agricultural fields, soil quality assessment, natural resource monitoring, soil erosion risk mapping, and solute transport in the vadose zone, requires more elaborate soil maps to depict spatial variability of soil properties (Brus and Noij, 2008; Chen et al., 2011; Corwin et al., 2010; Duffera et al., 2007). Moreover, with the technological development of GPS, remote and proximal sensing techniques, and computational advances of GIS, geostatistics, and data mining, high-resolution digital soil mapping (DSM) has almost replaced traditional maps and become a powerful approach to predicting continuous and quantitative soil properties with uncertainty (McBratney et al., 2003). The key concept behind the DSM is a comprehensive mathematical and statistical relationship between measured finite soil properties data and high-density and readily available environmental data, such as topography, digital elevation model (DEM), electromagnetic induction data, spectra data and other soil attributes (Taghizadeh-Mehrjardi et al., 2014). Target soil properties at new locations could be quantitatively estimated by the predictive model when ancillary environmental data is available at those locations (Scull et al., 2003).

An integral component of DSM process is a sound sampling design (SD) that provides a blueprint for collecting representative samples covering the whole field, thus obtaining a reliable input for establishing prediction model with environmental variables (Kidd et al., 2015a). SD is of great importance, since it affects the results of subsequent laboratory measurement and data analysis (de Zorzi et al., 2008). Total error can be divided into sampling error and analytical error, which account for more than 90% and less than 10%, respectively (Lame and Defize, 1993). Markert (2007) also illustrated that error caused by unrepresentative SD is much more than the error associated with sample preparation, instrument, or data analysis.

Sample size is an important component of a SD that has to be determined at first. In most cases, sample size is selected by available budget with the consideration of field work and laboratory analysis. On the other hand, if precision is the predominant factor, sample size will be chosen to cover the spatial variability which usually requires more samples. The balance between budget and acceptable accuracy determines the optimal sample size (Brungard and Boettinger, 2010). Vašát et al. (2012) compared the impacts of different numbers of soil samples on variogram

parameters and concluded that 50 samples are enough to prepare a reliable interpolation map of a 24 ha agricultural field (approximately 2 samples per ha). Sample locations is another essential component which is determined by different SDs. Sampling design is the selection procedure rather than an actual set of sample locations (Brus and de Gruijter, 1997). It has different purposes, such as searching for polluted site (Theocharopoulos et al., 2001); inference of population parameters (population mean and variance)(de Gruijter and ter Braak, 1990); or estimating the variogram (Lark, 2002). The main purpose of SDs in DSM is to provide reliable input for predictive models. Brus and de Gruijter (1997) discussed two approaches of soil sampling: design-based, which is mainly based on probability theory, and model-based, which comes from geostatistical analysis. However, these two fundamental approaches proposed here were mainly used for estimation of some statistical parameters. Most SDs in DSM are designed to provide a good spatial coverage of an area or a good coverage of variation of variables (Minasny and McBratney, 2006). Therefore, for purpose of spatial prediction of environmental covariates, another two categories are more proper: sampling in the geographical space or sampling in the feature space, which were proposed by Minasny and McBratney (2007).

A sampling design can be optimized in geographical space, feature space, or both (Hengl et al., 2003). Some studies argued that for calibration purpose, a good spatial coverage is not necessary in comparison with an appropriate cover of environmental variables (Minasny and McBratney, 2006). Nevertheless, more and more studies recommended obtaining a good spread in both feature space and geographical space. Hengl et al. (2003) recommended that an optimal SD should simultaneously represent the variation of soil properties in feature space and geographical space. They proposed an equal range stratification method where the range of the predictor variable is divided into equal-width clusters and then samples are randomly selected in each cluster according to given weights. Brus and Heuvelink (2007) also suggested that a good spread in feature space ensures an accurate estimation of regression coefficients, while reliable interpolation of sample data depends heavily on a good dispersion in geographical space. Walvoort et al. (2010) found that the estimation of spatial means of environmental variables became more accurate when the sample locations evenly spread in geographical space.

Our purpose in this review is to provide a detailed description including rationale, advantages and disadvantages of several commonly and widely used SDs in DSM. Furthermore, we hope to elaborately review recent studies (2010-2015) on how those SDs were used in DSM and compare

their performance and quality. Finally, we made discussions on the possible practical issues and solutions in SDs, summarized how those SDs were used in validating digital soil maps, and illustrated the applications and future work of SDs on three-dimensional digital soil mapping (3D-DSM).

2.2 Different sampling designs

A literature survey was conducted in Web of Science (Thomson Reuters) in order to thoroughly compile the recent publications (2010-2015) of case studies of DSM and SDs. Geoderma, Soil Science Society of America Journal, and European Journal of Soil Science were search with digital soil mapping as key words and with the period from 2010-2015. 146 papers popped out with such searching criteria, while 95 papers were finally selected after ruling out review papers and irrelevant papers. 31 papers used legacy dataset, generally, for large-scale digital soil mapping, such as mapping SOC in a continental-scale of Australia (Odgers et al., 2012), a national-scale mapping of SOC in Denmark (Adhikari et al., 2014), and 3D mapping of particle size distribution in Nigeria (Akpa et al., 2014). Since the historical dataset were collected for various soil survey, irregular sampling patterns were always displayed after merging diverse legacy datasets. Therefore, data selection and harmonization are necessary before mapping. As for the rest 64 papers, 14 papers did not describe which SD was used to conduct the soil survey and 8 papers showed ambiguously that sampling points were chosen to cover the variation or for convenience which could be regarded as purposive sampling (PS). Finally, among the 42 papers, simple random sampling (SRS), grid sampling (GS), cluster sampling (CS), transect sampling (TS), nested sampling (NS), spatial coverage sampling (SCS), stratified random sampling (StRS), Latin hypercube sampling (LHS) with its modification, and fuzzy-k means sampling (FKM) have been used for sample collection. A detailed introduction, development, advantages and disadvantages, as well as applications of these SDs and some other unmentioned SDs in DSM were demonstrated below.

2.2.1 Statistical and geometric sampling designs

The SRS, GS, CS, TS, NS, SCS, and StRS are classified as statistical and geometric SDs. Basically, these classical SDs are selected based on rigorous statistical inference or with the purposed of evenly spreading locations in geographical space (Royle and Nychka, 1998).

The SRS is the most basic probability sampling design in which each sample unit is selected randomly and independently. All potential units have the equal probabilities to be included and

thus allowing us to get unbiased estimates of mean and variance (Webster and Lark, 2013). The SRS is usually applied to relatively uniform and homogeneous fields and it is easy to implement (Fitzgerald, 2010; Zhang and Zhang, 2011). However, the SRS is generally regarded as an inefficient SD as it fails to utilize any available environmental or empirical information to reduce the sample size. It could not artificially place more points in suspicious points with higher spatial variability in a heterogeneous field. In addition, sampling points are easy to form clusters in this SD and may not guarantee a good spatial coverage (Fitzgerald, 2010). Therefore, only few papers used this SD for DSM: Evans and Hartemink (2014) and Adhikari and Hartemink (2015) used SRS for mapping subsoil red clay and topsoil SOC, respectively.

The GS provides us even coverage of geographical space and it is implemented by dividing the study area into regular grids and then selecting the nodes of grids as sample sites. Regular square grids and rectangular grids have been commonly used by various studies (Chaplot et al., 2010; Jonard et al., 2013; Piikki et al., 2013; Poggio et al., 2013; Rossel et al., 2010). Additionally, equilateral triangular grid was also used by Michot et al. (2013) and Malone et al. (2011) and it is regarded more efficient because the distance between the center of grid and sampling point is the lowest in a triangle (Webster and Lark, 2013). When conducting a reconnaissance survey for a field that has never been sampled before, a practical sampling approach might be the GS, supplemented with points at shorter distance (Brus and Heuvelink, 2007). However, the GS is accompanied with some statistical and applied constraints. It gives biased estimates of means and it faces practical constraints as the regular configuration is difficult to achieve in areas which are irregularly shaped or the domain that cannot be achieved (Brus et al., 2006).

The CS was developed with considerations of practical access issue especially in jungle or rain forest. It could reduce the travelling cost, effort, time, and sampling size without compromising the accuracy. In CS, several close sampling units consist of clusters that are selected randomly. Once a cluster is selected, all the sampling points within the cluster will be included (Brus et al., 2011). The CS could not spread sampling points evenly in the field, while it is recommended in rough region with spatial constraints. It has been adopted by Cambule et al. (2013) for DSM in poorly-accessible areas.

In TS, sampling points are usually arranged into a line and could be regarded as a cluster. (de Gruijter and Marsman, 1985). According to Thomas et al. (2012), sampling along topo-sequence transect is more efficient than random sampling sites in rugged and mountainous region by

maximizing the sampling rate and improving the efficiency. The TS has been used by Karunaratne et al. (2014) for mapping SOC fraction and by Liess et al. (2012) for mapping soil texture in mountainous area. In addition, it has been used as complementation to other SDs by Samyn et al. (2012) and Qin et al. (2012) to represent the relief feature. Furthermore, the fixed transect sampling interval could be modified to adopt variable sampling intervals (nested intervals), such as 0 m, 3 m, 30 m, 150 m, 500 m and 1,500 m (Thomas et al., 2012), to capture the variability in various distance ranges. This refers to the NS.

The NS is a hierarchical sampling that enables one to partition variance into contributions from different levels of a design (Webster and Lark, 2013). Spatially nested sampling design, with the same theoretical basis, is used to estimate the components of variance at two or more spatial scales. NS can be balanced or unbalanced, and generally unbalanced nested sampling is more flexible and efficient due to little influence on variance of lower stages (Webster and Lark, 2013). The nested intervals (0 m, 3 m, 30 m, 150 m, and 1500 m) chosen by Thomas et al. (2012) belonged to unbalanced nested sampling, while they recommended that 1,500 m sites should be abandoned on account of time-consuming access to such a large distance. The NS is usually not applied alone, but as a complement, added to other SDs to capture the variability in various distance ranges.

The SCS was introduced by (Walvoort et al., 2010) that aims to obtain even spatial coverage by minimizing the mean squared shortest distance (MSSD) by k-means clustering algorithm. The SCS is not widely used in DSM that only Kempen et al. (2015) has utilized it in mapping soil class in a national-scale. In addition, the SCS is similar to a spatially stratified SD proposed by Brus et al. (2006) in which they used geographical coordinates as objects and took x- and y- coordinates of the midpoint of the cells as classification variables and applied k-means clustering algorithm in geographical space. The centroids of the clusters were then chosen as sample points. Minasny and McBratney (2006) also used this method as stratified spatial sampling (StSS) design and created a good geographical coverage of the whole area.

In StRS, a region is equally or unequally divided into several strata based on prior environmental information and then a few samples are selected in a random manner within each stratum (Cochran, 1946). The sample size of each stratum is determined by the area of stratum. Stratification allows similar attributes and less variability within each stratum, while different attributes between strata. Various environmental variables have been used for stratification, such as DEM (Jafari et al., 2012), land use (Vasques et al., 2010), parent material (lithology) (Sun et al., 2012b), soil class

(Karunaratne et al., 2014), drainage class (Sun et al., 2012b), ECa (Koszinski et al., 2015), and solar insolation (Brown et al., 2012). In addition, equal squares could be also used for stratification which was also called SyRS and used by Kerry et al. (2012). Furthermore, stratification process was sometimes combined with other SDs to optimize the final pattern. For example, Sun et al. (2012b) stratified the study area based on parent material and drainage class, and then SCS was further used to select sample sites within each stratum. Cambule et al. (2013) combined stratification and cluster processes to select easily accessible sample sites. The StRS is an efficient SD which avoids clustering of sampling points, requires fewer samples to achieve the same accuracy, and obtains more even coverage within domains and in environmental variables.

Various previous studies have compared the efficiency of these SDs by considering its ability of reproducing the distribution of original covariates or the RMSE between measured and predicted soil properties. Wheeler et al. (2012) obtained a better soil carbon distribution of StRS over SRS as it reduced sampling variance by division of known sources of variation. Falk et al. (2011) compared SRS, GS, StRS, cLHS, and another spatially stratified sampling which is based on local spatial autocorrelation of auxiliary information and reported a lowest accuracy of SRS followed by GS and a generally better result of cLHS over StRS. Similarly, Thomas et al. (2012) compared SRS, GS and modified cLHS in a mountainous region and obtained the same rank as Falk et al. (2011).

However, these statistical and geometric SDs did not show any superiority compared to cLHS, FKM, and response surface sampling (RSS). Minasny and McBratney (2006) and Minasny and McBratney (2007) obtained slightly biased distribution of SRS in representing environmental variables compared to cLHS. In addition, despite a good spatial coverage obtained by StSS, it still cannot properly and accurately reproduce the original distribution of environmental variables. Worsham et al. (2012) showed the worst performance of SRS, followed by StRS, compared to cLHS in their research. Mulder et al. (2013) demonstrated that SRS and GS could not obtain a good coverage in feature space in comparison with constrained LHS. Corwin et al. (2010) compared StRS with RSS and argued that both SDs resulted in nearly equivalent salinity maps, but RSS exhibited a better range of ancillary data, hence an accurate calibration model.

Table 2.1 summarizes these basic SDs by considering six different factors. First, only SRS selects sampling locations randomly and independently, while some studies argued that it is not a necessary criterion for calibration purpose (de Gruijter et al., 2010). As for application condition,

SRS and GS could be performed in relatively uniform and homogeneous fields, and these are widely used in reconnaissance survey (Brus and Heuvelink, 2007). Available environmental information of the field is needed for StRS while CS is adopted when sampling is constrained by accessibility. TS is generally selected along topo-sequence and NS is applied to capture multi-scale variation. Furthermore, SCS provides the most even spatial coverage, followed by GS. StRS by stratification process, could also achieve good spatial coverage. SRS usually form clusters. CS, TS, and NS could not satisfy a good spatial coverage. Considering the simplicity in the design and implement stages, SRS, GS, and SCS are the simplest SDs without extra effort. StRS, TS, and NS is more complex as it collects and analyzes environmental information and spatial variation, while CS is the most complicated due to the access issue. As for accuracy and efficiency, StRS was reported with higher accuracy and widely used in previous studies. SRS was considered with lowest accuracy. However, the selection of SRS is a random process, its application might be restricted due to lack of robustness. GS and SCS are purely optimizing the pattern in geographical space and could be applied when environmental information is not available. CS, TS, and NS with specific purposes, could be applied in certain conditions.

Table 2.1 Comparison of different sampling approaches optimizing in geographical space

SDs	Independence[a]	Application condition	Spatial coverage	Simplicity	Accuracy	Efficiency
SRS	Yes	Homogeneous field	+	+++	+	+
GS	No	Homogeneous field	+++	+++	++	+++
CS	No	Access issue	+	+	+	++
TS	No	Generally topo-sequence	+	++	++	++
NS	No	Multi-scale variation	+	++	++	++
SCS	No	No	+++	+++	++	+++
StRS	No	environmental information	+++	++	+++	+++

a) independence: sampling locations are independently selected.

2.2.2 Geostatistical sampling

Geostatistical sampling optimizes the sampling pattern in geographical space by minimizing the average kriging variance (AKV) or maximum kriging variance (MaxKV) (Vašát et al., 2010). The spatial correlation and scale of variation of soil properties can be determined by variograms;

therefore, variograms can be used to guide SDs (Kerry and Oliver, 2004). This can be done by two methods: choose sampling interval of less than half the range of spatial dependence; or determine sampling interval by optimizing kriging equations to achieve a tolerable error (Kerry and Oliver, 2004). Knowing the model and spatial structure of the residues of the model are two prerequisites for geostatistical sampling. Minasny and McBratney (2007) used simulated annealing (SA) to optimize sample pattern for universal kriging of environmental variables with known variogram of residues and the assumption of linear predictive model. However, the residual variograms and predictive model are usually unavailable for an unknown area. In the reviewed papers, none of them utilized this design, whereas it can be helpful to improve the sample patterns on a previously investigated area.

2.2.3 Latin hypercube sampling (LHS)

As for calibration exercise in DSM, ancillary environmental covariates, containing some relationship (such as linear relationship (Brus and Heuvelink, 2007)) with soil properties of interest, can be used to predict soil properties and guide soil SD (Minasny and McBratney, 2006). The stronger the relationships, the more accurate and reliable SD and predictive model could be obtained (Brus et al., 2006). However, a good coverage in geographical space might not guarantee a good coverage in all the environmental covariates. Therefore, more attention has been focused on sampling in feature space which is a virtual space describing the distribution or range of environmental covariates (Lillesand and Kiefer, 1994). Once a good coverage of values of environmental covariates is obtained by a SD, a full representation of expected soil properties is enhanced, thus reducing the uncertainty and errors caused by extrapolation (Minasny and McBratney, 2007).

Latin hypercube sampling (LHS) is a maximally stratified random sampling procedure that guarantees a full coverage of multivariate distribution. It was originally developed to effectively select a set of values of input variables for computer models (McKay et al., 1979).

The LHS works as follows (Minasny and McBratney, 2006):

- In order to select *m* samples from *k* variables, firstly we divide the cumulative distribution of each variable into *m* intervals with equal probability.
- A sample is randomly taken at each interval, so that *m* values will be selected for each variable.
- Then the *m* values obtained from *k* variables will be matched with each other in a random manner.

- Finally, we get *m* values which cover the full distribution of all variables.

In order to avoid nonexistent points caused by randomly matching multivariate distribution, Minasny and McBratney (2006) proposed conditioned Latin hypercube sampling (cLHS) by adding a search algorithm based on heuristic rules and an SA process (van Groenigen et al., 1999), hence obtaining a true or approximate Latin hypercube of multivariates in feature space.

Up until now, LHS or cLHS has been widely applied on DSM and proved to be efficient to characterize soil properties. Table 2.2 summarized the previous studies on DSM using LHS or its modified forms. Various covariates have been used to develop LHS, while relief including topography information is the most common one. Moreover, on the basis of LHS and cLHS, more optimization criterions were added to the original algorithm to achieve specific purpose. In order to place more sampling points on the edges of the distribution, the D-optimality Latin hypercube sampling (DLHS) was proposed by de Gruijter et al. (2010) by adding a D-optimality criterion (Laycock and Lopez-Fidalgo, 2007). However, DLHS did not exhibit any superiority compared to cLHS unless the calibration model is a linear model (Louis et al., 2014; Minasny and McBratney, 2010). For the purpose of quantifying the cost of reaching every point and enhancing the sampling efficiency, Roudier et al. (2012), Mulder et al. (2013), and Clifford et al. (2014) utilized an operational algorithm to produce a cost layer by incorporating terrain attributes, land cover, and travel time and added this cost layer in cLHS as an environmental covariate. However, this method cannot achieve the same accuracy as standard cLHS due to an under-sampled region in the rough area, thus a balance between accuracy and budget should be considered when using this method. In addition, weighted conditioned Latin hypercube sampling (wecLHS) with extremes was developed by adding a weighing scheme for the purpose of reducing the noise of signal and setting the extreme values (min and max) of all sensors (Schmidt et al., 2014). An effective covariates data reduction method using iterative principal component analysis (iPCA) was combined with cLHS by Levi and Rasmussen (2014).

Previous studies have compared cLHS with other SDs and pointed out that cLHS is an effective and better way to reflect the original distribution of soil properties. Minasny and McBratney (2007) compared cLHS to SRS, StRS, and sampling along the PCs and indicated superiority of cLHS over other SDs. Similarly, Worsham et al. (2012) compared cLHS with SRS and StRS and concluded a good coverage of target soil carbon content of cLHS. Mulder et al. (2013) applied LHS with constraints and compared it with SRS and SyRS. The authors demonstrated that cLHS is superior

Table 2.2 Summary of previous DSM studies (2010-2015) using LHS with its modification forms

References	SDs	Environmental covariates							Spatial extent	Profile size	Output maps		Predictive model	Validation		
		s	c	o	r	p	a	n			S_a	S_c		Split	CV	ID
Taghizadeh-Mehrjardi et al. (2015)	cLHS								3,000ha	217		×	LR,ANN,SVM, KNN,RF,DT	×		
Thomas et al. (2015)	fLHS	×	×		×				155,000km^2	1951	×	×	Rulefit3	BS	×	SRS
Brungard et al. (2015)*	cLHS	×			×				190km^2	103		×	LDA,MNLR,K		×	
			×	×	×				300km^2	300		×	NN,CT,RF,NN,		×	
					×				296km^2	57		×	SVM		×	
Rad et al. (2014)	cLHS	×		×	×				85,000ha	99		×	RF			
Levi and Rasmussen (2014)	PCA+cLHS	×	×	×	×	×			6250ha	52	×		OK,RK		×	
van Zijl et al. (2014)	cLHS+TS				×				10,970ha	206		×	SoLIM	×		
Taghizadeh-Mehrjardi et al. (2014)*	cLHS	×	×	×	×	×	×		72,000ha	173	×		RT	×		
Lacoste et al. (2014)	cLHS	×		×	×				10km^2	70	×		Cubist			StRS,TS
Odgers et al. (2011a) and Odgers et al. (2011b)	LHS			×	×	×			-	262		×	FKM,RK			
Silva et al. (2015)	cLHS				×				1.6ha	12	×	×	IDW			×
	ccLHS				×			×	1.6ha	12	×	×	IDW			×
de Brogniez et al. (2015)	cLHS				×				Europe	200,000	×		GAM	×		

Sampling designs: cLHS- conditioned Latin hypercube sampling; ccLHS- cost-constrained Latin hypercube sampling; fLHS- flexible Latin hypercube sampling; PCA+cLHS- principle component analysis for covariates used in cLHS; TS- transect sampling.

Environmental covariates (used in soil sampling design, not in the digital soil mapping process) : s- soil; c- climate; o- organism; r- relief; p- parent material; a- time; n- spatial location.

Output maps: S_a- soil attributes; S_c- Soil classes.

Predictive models: ANN- artificial neural network; CT- classification tree; DT- decision tree; FKM (FCM)- fuzzy k-means clustering ; GAM- general additive model; IDW- inverse distance weighting; KNN- K-nearest neighbor; LDA- linear discriminant analysis; LR- logistic regression; MNLR- multinomial linear regression; NN- neural network; OK- ordinary kriging; RF- random forest; RK- regression kriging; RT- regression trees; SoLIM- Soil Land Inference Model; SVM- support vector machine.

Validation: Split- split dataset into calibration dataset and validation dataset, including bootstrapping method (BS); CV- cross-validation; ID- independent dataset; SRS- random sampling; StRS- stratified random sampling; TS- transect sampling.

to other statistical SDs, due to a good coverage of feature space and target soil properties. In addition, it also outperforms other SDs which optimize the sampling patter in feature space. Ramirez-Lopez et al. (2014) showed superiority of cLHS to KSS and FCMS. Schmidt et al. (2014) compared wecLHS with extremes to FKMS and RSS and demonstrated four advantages of wecLHS including i) a better coverage of covariate space, ii) preservation of correlation between sensor data and target soil properties, iii) inclusion of extreme values of sensor data and iv) weighing scheme that helps to pay more attention to the strongest soil response.

However, some papers also pointed out some drawbacks of LHS. Thomas et al. (2012) demonstrated that cLHS inhibits the interaction with surveyor's landscape experience, which can be overcome by fuzzy clustering. Falk et al. (2011) concluded that cLHS is less accurate to capture high parameter levels in comparison to spatially stratified sampling. de Gruijter et al. (2010) put forward two drawbacks of LHS: i) LHS is random sampling procedure in feature space, while for calibration purpose, there is no need to select sample points randomly; and ii) LHS is not prone to choose sampling points at or near extreme values which is expected in calibration sample set. Nonetheless, with more criteria developed to satisfy specific purpose in particular mapping process, LHS has been getting optimized to overcome such kind of disadvantages. In general, it is the most widely used SD due to its full coverage of all the environmental covariates and its feasibility to add more optimization criteria to achieve specific purpose.

2.2.4 Fuzzy k-means sampling

K-means clustering algorithm is a classification method which aims to create clusters within which the objects share the similar attributes while different from other objects in different clusters (McBratney and Degruijter, 1992). These clusters are optimized by least square criterion, which was initially developed by Hartigan (1975). However, considering overlapping attributes of objects in different clusters and improper sharp boundaries between clusters, Bezdek (1981) extended the algorithm with degree of fuzziness and created more continuous clusters (fuzzy sets), in which the objects have memberships varying between 0 and 1. In order to cover the extremes of corners, an amendment was produced by adding an extragrade class (McBratney and Degruijter, 1992). These are the rationale for development of FKMS design.

The FKMS also marked as FCMS, was proposed to choose sampling point by using the fuzzy classification and membership criteria (de Gruijter et al., 2010). The procedure works as follows: first take the environmental data in state space as objects and apply k-means clustering analysis to

create k fuzzy subsets. And then choose the objects with the largest membership in the subsets as sample points, and search for corresponding geographical position to obtain the sampling pattern. Fuzzy k-means sampling with extragrades (FKME) can represent the extremes in corners and improve the predictive ability of the traditional FKMS. Geographical coordinates can also be used as objects to conduct fuzzy k-means clustering algorithm, while this is a method to optimize sampling pattern in geographical space (Brus et al., 2006).

Fuzzy k-means is often used for preparing soil classification maps. For example, Chapron (2011) conducted a classification of soil and vegetation using fuzzy k-means for precision agriculture, and Burrough et al. (2000) successfully applied fuzzy k-means classification of landscapes from DEMs. Triantafilis et al. (2012) used Fuzzy k-means based on total counts and three radioelements (K, U, and Th) and produced 11 fuzzy classes. As for FKMS design, Ramirez-Lopez et al. (2014) compared FKMS with cLHS and Kennard-Stone sampling (KSS) and concluded that FKMS and cLHS can better reflect the original vis-NIR spectra distribution than KSS. While Schmidt et al. (2014) concluded that wecLHS is better than FKMS as FKMS exhibits a clustered distribution and cannot obtain extremes. Kidd et al. (2015a) used fuzzy k-means to classify the field and then randomly chose sample sites within each stratum in reserve for original cLHS in case of some inaccessible sample locations.

By applying fuzzy classification process, FKMS is able to cover all the meaningful classes and choose the representative point of each class (Burrough et al., 2000). However, FKMS fails to recognize the categorical variables, so that un-ordered categorical variables, such as soil type cannot be used to guide sampling design (Kidd et al., 2015a). They also indicated that potential weakness boils down to an unreasonable partition of sampling points in each stratum that smaller clusters in FKMS design are assigned the same sample numbers as the larger clusters. However, they rationalized this partition that smaller clusters with higher variability requires smaller sampling interval, thus the same numbers as the larger clusters. Further research must be focused on cluster size and sample density in each clusters.

2.2.5 Response surface sampling

Response surface sampling (RSS) applies a response surface design with the goal of optimizing the estimation of linear regression model parameters, as well as achieving approximate spatially independent regression residuals (Schmidt et al., 2014). The RSS is basically used to measure the soil salinity (ECe) by conductivity information (ECa) (Lesch et al., 1995). Additionally, many crop

or soil variables (e.g., crop biomass, soil texture, soil salinity) can be measured by this approach (Corwin and Lesch, 2005).

The procedure of this sampling design is as follows (Fitzgerald et al., 2006; Lesch, 2005):

- Transform and decorrelate signal data using principle component analysis (PCA). Outliers are removed in this process.
- Afterwards, apply traditional RSS design, like second-order central composite sampling design, and generate original sampling design.
- Finally, by optimization criterion and iterative algorithm to spread out the sampling points to obtain optimal sampling design.

The whole process can be implemented by the ECe Sampling, Assessment, and Prediction-response surface sampling design (ESAP-RSSD) software program, which is designed to create optimal sampling designs from bulk soil conductivity survey information (Lesch et al., 2000).

Like other SDs, several advantages and disadvantages were reported in previous studies. The RSS can select a minimum set of calibration samples and the sample locations are unambiguous in comparison with SRS and co-kriging. At the same time, the number of samples are restricted by the computer program, since only 6, 12 or 20 samples can be chosen (Corwin and Lesch, 2005; Fitzgerald, 2010). The spatial separation of sampling positions and value separation of data can be achieved which help to sample more extreme values and reduce the probabilities of extrapolation errors (Fitzgerald, 2010). This approach can naturally analyze the remotely-sensed data, thus optimizing the sampling process in an efficient manner (Lesch, 2005).

2.2.6 Kennard-Stone sampling

Kennard-Stone sampling (KSS) (Kennard and Stone, 1969), originally called uniform mapping algorithm, is a deterministic sequential method that selects sampling points uniformly distributed in the state space (Ramirez-Lopez et al., 2014).

The sampling procedure is as follows (Dieterle, 2003):

- When choosing n samples from a set of N samples, the algorithm starts from finding two samples that are farthest apart from each other, and keep the two samples in the calibration set.
- Then repeat the procedure until the expected number of calibration set.

Ramirez-Lopez et al. (2014) compared KSS with FKMS and cLHS, and noticed that KSS tended to select a wider range of soil attribute values due to the extreme samples selected by KSS algorithm over others.

2.3 Discussions

2.3.1 Operational challenges and solutions

While a predefined SD can provide an easy and efficient way to carry out field work, it is not always worry free and devoid of challenges. In many cases, when a predefined sampling scheme is applied in a real world, it might become impractical due to some access issues and operational challenges. Kidd et al. (2015a) summarized possible access issues inherent in the real world, including physical or consensual access, cropping, disturbance, infrastructure, livestock, stone, terrain, biosecurity and conservation. This calls for methods that are flexible and compatible to implement and can increase the operational efficiency without compromising the sampling accuracy. They proposed a method that manually re-locate inaccessible locations by easily accessible locations which were randomly selected within clusters of the same type determined by fuzzy k-means classification technique, while being consistent with the same number as original sampling design.

Additionally, Roudier et al. (2012) quantified the cost of accessing sampling points from roads by terrain and land cover attributes and incorporated the new access layer into cLHS, thus obtaining a sampling scheme that was easier to implement. However, the accuracy was reduced in the cost-effective sampling design due to under sampling in slopes and rainfall regions. Moreover, they provided two resolutions to improve the scheme: i) the weight between environmental covariates and access layer should be carefully balanced; ii) the access layer might be calculated in an sequential way from point to point rather than evaluating the distance between roads and sampling points. On this basis, Clifford et al. (2014) developed a flexible LHS by adding two optimization criteria for spatial spread and ease of access to original algorithm. This design aimed at evenly covering the feature space and geographical space, focusing on more easily reached sites, and providing alternative sites.

Thomas et al. (2012) encountered access issues due to the rugged terrain and remote area of mountainous region, and they combined cLHS with topo-sequence transects with nested intervals. The design obtained a good landscape coverage, increased efficiency and captured variability of various distance ranges. Cambule et al. (2013) proposed a methodology that soil properties in poorly-accessible regions could be predicted by the model built on accessible regions when there are similarities of environmental variables and soil properties of interest between accessible and inaccessible regions. Even though the performance of a case study was not as good as expected, it

was still a good method as the models showed similar predictive quality in both accessible and poorly-accessible regions.

In addition, van Groenigen and Stein (1998) used spatial simulated annealing sampling (SSAS) scheme, combined with considerations on physical sampling constraints and delineations, and existing information. This proved to be a good solution to such a situation when buildings and water areas in urban region can't be assessed during sampling process. This SSAS optimization method was also used by Scudiero et al. (2011) in a salt-contaminated costal farmland, by taking into account physical sampling constraints, filed shape and existing ECa maps.

2.3.2 Sampling designs for validating DSM

In addition to calibrating digital soil maps, another purpose of sampling is to estimate the quality of the maps produced (de Gruijter and Marsman, 1985). The accuracy simply obtained by comparing the predicted values and measured values in the calibration dataset is regarded as internal accuracy which always overestimates the actual accuracy (Chatfield, 1995). Therefore, an independent testing (or validation) dataset which is not used in the calibration process should be searched for a more reliable quality estimates. This is validation process (Williams, 1996). The accuracy obtained is considered as test accuracy or external accuracy. There are three common methods for obtaining validation dataset.

Data-splitting is one method to extract subsamples from a calibration dataset and the sub samples reserve as a validation dataset. For example, the simplest and most common way to select a validation dataset is just randomly setting aside a small portion (20% or 30%) of the calibration dataset, but this partition lacks a statistical basis (McBratney et al., 2003). Data-splitting has been used by Ramirez-Lopez et al. (2014) for validating digital soil maps, and by Veronesi et al. (2012) to validate 3D soil compaction map. In addition, another splitting method bootstrapping is also proposed to select subset.

Bootstrapping is based on sampling with replacement to determine a calibration dataset. The procedure is (Molinaro et al., 2005):

- Firstly, select a sample from *t* samples with replacement.
- And then repeat *t* times.
- At the end, about 63.2% samples will be selected as a calibration dataset and the 36.8% samples left will be a validation dataset.

Cross-validation (CV) is another effective method where validation is repeated several times, and an average error is obtained eventually. It works as follows (Dieterle, 2003):

- For a *n*-fold cross-validation, the dataset is partitioned into *n* equal parts.
- Firstly, the first part is used as a validation dataset, and the rest data are used as a calibration dataset.
- Then the second part is used as a validation dataset, and the rest data are for calibration.
- Repeat the procedure for *n* times so that each part will be used as a validation dataset once and we will obtain *n* validation results.
- Average the validation results as the ultimate result.

Leave-one-out cross-validation (LOOCV) is the most common form when *n* equals the sample size so that each sample will be used as a validation dataset (Dieterle, 2003). Leave-one-out cross-validation is usually considered as the best choice for small datasets due to limited budget and time.

Schmidt et al. (2014) compared 10-fold CV, LOOCV, bootstrapping and 632 bootstrapping which is an optimization of original bootstrapping method. They concluded that LOOCV and 632 bootstrapping are good validation method over others. Additionally, Mueller et al. (2004c) compared the performance of an independent dataset and CV and concluded that the independent dataset outperformed CV. One problem about the data-splitting method is that calibration dataset is usually selected according to purposive SDs, the subsamples extracted from calibration dataset may not be unbiased and the partition process is often unclear (Brus et al., 2011).

Furthermore, collection of an independent dataset following a SD as the third method is regarded as the most credible way to select validation dataset. All of the aforementioned SDs can be used to select validation dataset, whereas Knotters and Brus (2013) recommended probability sampling for map validation due to the unnecessary model assumption. Collection of validation samples is usually concurrently conducted with calibration samples to save time, money and effort (Kidd et al., 2015a). A detailed comparison and summary of four basic probability SDs (SRS, StRS, SyRS, and CS) was made by Stehman (1999), Stehman and Czaplewski (1998), and Brus et al. (2011) by considering simplicity, cost, precision, spatial coverage, and estimation error. The results of comparison are similar as of our comparison presented in section 2.2.1. Brus et al. (2011) argued that StRS is an optimal option considering all the pros and cons. Mueller et al. (2004b) utilized SRS to select independent validation dataset to estimate the quality of soil property maps and

concluded that sampling intensity might be adjusted to improve the prediction quality. Mueller et al. (2004a) used TS to choose a validation dataset which guarantees that the calibration dataset and validation dataset were collected in perpendicular direction, and then validation was performed for soil electrical conductivity maps. Stehman (1999) suggested that every probability sampling is available for validation, while specific sampling design should be selected according to project objectives.

In addition, more complex SDs were also used in validating digital soil maps. Fuzzy k-means clustering method was used to select a validation dataset by Kidd et al. (2015a). Even though this is not a probability sampling design, in the real world, any SD cannot satisfy the equal-probability criteria (Kidd et al., 2015a). Moreover, the prediction error was quantified at each sampling point in the validation dataset by this method and the mean prediction error can be used as an overall prediction model error (Laslett et al., 1997). Kidd et al. (2012) used FKMS to choose six clusters and 10 random samples within each cluster for a validation dataset, and it proved to be a good validation design.

The review of previous literatures showed a similar percentage of using the three methods (splitting, CV, and independent validation dataset). In summary, CV (especially LOOCV) is a rapid and inexpensive splitting method to select validation dataset from full dataset when budget is constrained (Mueller et al., 2004b). However, using an independent dataset to validate digital soil property maps is more recommended. In addition, probability sampling designs without model assumptions are highly recommended in comparison to model-based sampling designs. Four basic sampling design with specific advantages and disadvantages can be used for validating, while the selection of those sampling designs is dependent on project objectives and available information. Moreover, model-based sampling designs have also been used for validation, while its efficiency needs further consideration. Up until now, there are limited papers about the comparison of different validation methods and their contribution to the whole mapping procedure. However, with the increase of digital soil maps, there is an urgent need to search for effective validation method.

2.3.3 Sampling designs in 3D digital soil mapping

With the flourish of three-dimensional digital soil mapping (3D-DSM), SDs provide instructions for collecting not only top soil samples but also the whole soil profiles. Various SDs haven been used for sampling soil profiles for 3D-DSM. For example, Taghizadeh-Mehrjardi et al. (2014)

used cLHS based on DEM and geomorphologic units to collect samples for 3D mapping of soil salinity; Lacoste et al. (2014) also used cLHS based on elevation, wetness index, gamma radiation of potassium, and grassland frequency for 3D mapping of SOC; Michot et al. (2013) and Malone et al. (2011) used triangular GS for 3D mapping of soil salinity, SOC and available water capacity (AWC), respectively; Vasques et al. (2010) used StRS based on soil order and land use to collect samples down to 180 cm for mapping soil carbon; ECa and DEM were used for StRS by Koszinski et al. (2015) to collect soil cores to 2 m in depth for mapping SOC. However, the sampling patterns created on the soil surface might not keep the same in deep soil due to various depths of different profiles. Namely, the GS might lose some points in deep soil, thus destroying the even pattern of SDs and forming under-sampled region. In addition, the SDs developed based on landscape or other environmental variables which are mainly collected in soil surface might not cover the actual variation in deep soil. On the contrary, parent material, terrain attributes, or other gamma ray radiometric attributes which could reflect the deep soil attributes might be superior to guide a SD. Likewise, sampling in feature space is optimal than classical statistical and geometric sampling due to the flexibility and better cover of attributes. While more work needs to be done to test the effectiveness of different SDs on 3D-DSM.

2.4 Conclusions

We have reviewed various SDs for calibrating digital soil maps. Generally, sampling in feature space is better than statistical and geometric sampling that optimize the sample pattern in geographical space. For sure, SDs that can simultaneously satisfy the variation in geographical space, feature space, and target soil properties are searched and preferred. The cLHS and its optimization format are the most commonly used and highly recommended methods. The StRS and SCS are most efficient methods that optimize sample patterns in geographical space. Nevertheless, SRS and GS are still commonly used and easily conducted methods when environmental information is not available for an unknown field. When considering access issues, CS and cLHS with cost function are generally used sampling method. TS could be applied in mountainous region along topo-sequence. NS works as a complementary design for understanding multi-scale variation. The FKMS, as a feature space stratification method, is also widely used for selecting sampling points to cover all the classes of environmental variables. The RSS that can concurrently optimize sampling pattern in geographical space and feature space is also recommended. The selection of a SD for specific study mainly is determined by the available

information one have and the particular target one wants to achieve. In addition, SDs for choosing validation dataset are also reviewed. Compared to the subsampling from full dataset, an independent dataset will result in a more reliable estimate of map quality. Moreover, owing to the free assumption of model, probability sampling is preferred for selecting sample locations. However, more work is needed to verify the performance of SDs on validating digital soil maps.

Sampling design is not an independent part of an overall DSM process. Its application should be incorporated with subsequent model selection and final validation stage. Therefore, rather than simply considering which sampling design is better, we'd better consider a complete mapping process and the best combination of sampling design, predictive model and model validation. Such kind of work has been done by Schmidt et al. (2014) who found that a combination of LHS and random forest regression is optimal. With the development of three-dimensional digital soil mapping to simultaneously quantify the horizontal and vertical variability, a single sampling design has been rarely assessed for its ability to capture the variability for multiple layers. Therefore, more works need to be done for applying and testing these sampling designs for 3D-DSM. In a changing world, there are more challenges for DSM, in consideration of sustainability, vulnerability, adaptability, and risk-assessment of soil-ecosystems across spatial and temporal scales (Grunwald et al., 2012).

Based on the literature review, six sampling designs were chosen for this thesis: 1) Simple random sampling (SRS). Although SRS was reported to have low efficiency, it was chosen as it is the most basic probability sampling with unbiased selection of sample locations; 2) Grid sampling (GS). GS is commonly used in DSM to obtain a relatively even coverage of the study area, especially for a reconnaissance survey without any available environmental information; 3) Grid random sampling (GRS) also known as systematic random sampling. GRS was chosen in order to obtain a good spatial coverage on the basis of GS and simultaneously make the GS more flexible; 4) Stratified random sampling (StRS). StRS was another widely used sampling design in DSM. It divided the field into strata base on available environmental information rather than regular grid in GS and substantially increased the efficiency of sampling design; 5) Transect sampling (TS). TS was usually used as topo-sequence transects and combined with other sampling designs. TS was identified with unequally (nested) distributed sample locations in this study; and 6) conditioned Latin hypercube sampling (cLHS). The previous five sampling designs were mainly identified and optimized in geographical space. In order to make the results more comparable and

persuasive, cLHS that mainly optimizes sample patterns in feature space was selected. Finally, six sampling designs with 45 sample locations were compared for 3D-DSM in the subsequent studies.

2.5 Digital soil mapping techniques

Lots of studies have been using two-dimensional mapping methods, either by interpolation techniques based on geostatistics theory such as ordinary kriging (Evans and Hartemink, 2014), universal kriging (Li et al., 2015a), block kriging (Vasques et al., 2010), and regression kriging (Ballabio et al., 2012), or data mining techniques such as classification and regression trees (Brown et al., 2012), support vector machine, artificial neural networks, and random forest (Taghizadeh-Mehrjardi et al., 2015). Only a few studies have discovered the 3D-DSM methods mainly using a combination of profile depth functions and 2D interpolation methods. For example, the equal-area quadratic spline function has been widely used with 2D mapping techniques in 3D-DSM of available water capacity (Malone et al., 2009), soil organic matter (Liu et al., 2013), soil texture (Adhikari et al., 2013), soil organic carbon (Lacoste et al., 2014), soil salinity (Taghizadeh-Mehrjardi et al., 2014), and cation exchange capacity (Taghizadeh-Mehrjardi, 2016). Polynomial function combined with ordinary kriging was used for 3D mapping of soil compaction (Veronesi et al., 2012). A linear function with a power function was used as profile depth function for 3D mapping of soil organic matter (Liu et al., 2016). Although these studies showed 3D structure in the final maps, the mapping process separately considered the horizontal and the vertical variabilities that did not make full use of the 3D spatial relationship. Another method for 3D-DSM is using 3D geostatistics, such as universal kriging for 3D mapping of soil texture (Veronesi, 2012). 3D regression kriging which integrates the regression techniques with interpolation methods was proposed as a simple, effective, and accurate method for 3D mapping of global soil grids at 1km resolution (Hengl et al., 2014) and Africa soil grids at 250m resolution (Hengl et al., 2015), respectively. It also allows for the adoption of multiple regression techniques, such as generalized linear model (Hengl et al., 2014) and random forest which was reported to greatly improve the prediction (Hengl et al., 2015). A national-scale 3D mapping of soil organic carbon also adopted the 3D regression kriging with multiple regression techniques (Mulder et al., 2016). However, 3D RK has not been widely assessed and used at a small-scale for various soil properties and regression methods. Therefore, in this thesis, 3D regression kriging was selected with three regression methods including generalized linear model, regression tree, and random forest for 3D-DSM of a set of physical and chemical soil properties.

PREFACE TO CHAPTER 3

Chapter 2 reviewed a set of commonly used sampling designs in digital soil mapping for their theory, development, advantages, disadvantages, application, and comparison. In addition, a short review of digital soil mapping methods, especially 3D digital soil mapping methods was presented. This greatly assisted in choosing the sampling designs in this study and conducting the DSM project. At the end of Chapter 2, it explicitly explained the reason for choosing six different sampling designs and 3D regression kriging with three regression techniques. Based on the background and principles discussed in Chapter 1 and Chapter 2, the actual 3D-DSM work was conducted and the results were reported for multiple soil properties in Chapter 4. In addition, visible near infrared (vis-NIR) spectroscopy is a popular technique to predict various soil properties and prepare for DSM. While it has been rarely used *in-situ* and in multiple soil depths. Therefore, the spectral models were developed for multiple soil properties and tested based on *in-situ* spectra data in Chapter 3 which is a preparation for the 3D-DSM in Chapter 4. A field experiment was implemented in a small agricultural field in McGill University. 19 soil properties were measured including VWC, GWC, BD, SOM, soil pH, EC, sand content, silt content, clay content, available P, K, Na, Ca, Mg, Zn, Mn, Al, Fe, and CEC due to the available techniques and high variability of these soil properties. Partial least square regression (PLSR) models have been used to calibrate the relationship between soil properties and vis-NIR spectra.

Chapter 3 has been written as a research paper format and will be submitted to *Soil Science Society of America Journal* (impact factor: 1.75). The detailed information for authors is shown below:

Order of authors: Yakun Zhang[a], Asim Biswas[a,*], Wenjun Ji[b], & Viacheslav I. Adamchuk[b].

The author of this thesis took charge of experimental design, field experiment, laboratory measurement, data analysis and interpretation, and manuscript preparation. Prof. Biswas as thesis supervisor was completely involved in this study and greatly assisted in the experimental design, technical support, scientific advice, and review and edition of the manuscript. Wenjun helped with the vis-NIR spectra data preparation and advised on the spectral model development. Prof. Adamchuk as thesis committee provided much support on the technical facilities, environmental variables collection and compilation, and data analysis.

CHAPTER 3

Depth specific prediction of soil properties *in-situ* using vis-NIR spectroscopy

Abstract:

Visible-near infrared (Vis-NIR) spectroscopy has been used to efficiently and accurately predict various soil properties and prepare for digital soil mapping. However, challenges exist for *in-situ* vis-NIR spectra collection due to the interference from soil moisture. In addition, with the development of three-dimensional digital soil mapping, the performance of vis-NIR spectroscopy on multiple soil depths gained more attention. Therefore, this paper aims to test the predictive ability of vis-NIR spectra on various physical and chemical soil properties in the field condition and its performance in whole soil profiles down to 1-m depth. Vis-NIR spectra (400-2200nm) were continuously collected *in-situ* at 32 locations from 0 to 1-m maximum depth in an agricultural field, Macdonald farm, McGill University. Soil cores were sampled at the same locations and sectioned at every 10 cm intervals. A total of 251 samples were measured for various soil properties in the laboratory, including volumetric water content (VWC), gravimetric water content (GWC), bulk density (BD), soil organic matter (SOM), soil pH, electrical conductivity (EC), sand content, silt content, clay content, available phosphorus (P), available soil cations (potassium (K), sodium (Na), calcium (Ca), magnesium (Mg), zinc (Zn), manganese (Mn), aluminum (Al), iron (Fe)), and cation exchange capacity (CEC). Partial least square regression (PLSR) models were developed to calibrate vis-NIR spectra against laboratory measured soil properties and validated by leave-one-out cross validation. The results showed that vis-NIR spectra could be used to predict soil organic matter and water-related soil properties with high accuracy, while other soil properties with some positive or negative relationship with SOM and soil water could also be fairly predicted. In addition, there is no direct influence on prediction by multiple depths, whereas it greatly affected the actual values of soil properties and hence corresponding predictions.

3.1 Introduction

Precision agriculture, with the concern of site-specific management of fertilizers and environmental sustainability, requires a thorough understanding of soil spatial variability (Lake et al., 1997). Intensive sample collection and conventional laboratory analysis of soil properties are expensive, time-consuming, and laborious (Viscarra Rossel and McBratney, 1998). As an alternative, proximal soil sensing involving various electrical, electromagnetic, optical,

radiometric, mechanical, electrochemical sensors has been used worldwide to rapidly obtain soil information at fine-scale (Adamchuk and Viscarra Rossel, 2010).

Vis-near infrared (vis-NIR) diffuse reflectance spectroscopy with a wavelength range of 350-2500 nm is a widely used technique as it can simultaneously predict multiple soil properties with a single scan (Stenberg and Viscarra Rossel, 2010). The overtones and combinations of fundamental vibrations of molecular bonds (e.g. OH-, CH-, NH-, and CO-) in the mid-infrared (MIR) region are detected in the vis-NIR region. This further generates different characteristic curves of spectra according to different constituents and concentrations of soil samples which could be used for assessment of soil properties (Stenberg et al., 2010). Many studies have investigated the prediction of diverse soil properties by vis-NIR spectroscopy. Soil organic matter (SOM) and organic carbon (OC) were amongst the most popular attributes and generally reported with ideal prediction (McCarty et al., 2002; Wang et al., 2015). Prediction of attributes related to soil pH (pH_{Ca}, pH_w, and lime requirement (LR)) was observed with moderate results (Viscarra Rossel et al., 2006c). Soil texture (sand, silt, and clay) and mineralogy which are fundamental constituents of soil were accurately predicted by NIR spectra (Vendrame et al., 2012). While poor results were illustrated on macronutrients (nitrogen (N), phosphorus (P), and potassium (K)) (Wang et al., 2015). Micronutrients (Copper (Cu), Manganese (Mn), iron (Fe), and Boron (B)) showed moderate to poor prediction (Terra et al., 2015). Additionally, Shepherd and Walsh (2002) demonstrated a good prediction with $R^2 > 0.75$ for exchangeable calcium (Ca), magnesium (Mg), and effective cation-exchange capacity (ECEC). However, majority of the current studies worked on air-dried and ground soil samples.

Measuring spectra in the field condition faces quite a bit of environmental and operational challenges. Interference from soil moisture and its high variability in the field greatly reduce the applicability and accuracy of vis-NIR spectra (Wang et al., 2016). In addition, contact issue between soil and probe and contamination from plant and crop residues also create noise in spectra (Stenberg and Viscarra Rossel, 2010). However, despite these difficulties, *in-situ* measurement of spectra has attracted much attention as it substantially simplifies the sample preparation procedures e.g. air dry and grinding. Few previous studies have examined the feasibility of *in-situ* measurement of vis-NIR spectra. For example, Daniel et al. (2003) used *in-situ* vis-NIR spectra with a range of 400-1050 nm to predict SOM, P, and K and demonstrated slightly worse results compared to laboratory measured spectra. Maleki et al. (2006) used vis-NIR spectra with range of

305-1710 nm on fresh soil to predict P and obtained good accuracy with $R^2 > 0.70$. Waiser et al. (2007) compared the *in-situ* measurement of clay content with measurements from air-dried soil, air-dried and ground soil, and smeared soil cores at the field moisture condition. They concluded that it is acceptable to measure clay content *in-situ* with various water contents. Morgan et al. (2009) observed slightly worse prediction results of OC and inorganic carbon (IC) from field moist samples compared to air-dried samples. Al-Asadi and Mouazen (2014) used vis-NIR spectra and partial least square regression (PLSR) to predict gravimetric water content (GWC) and obtained good results with R^2 of 0.91. Ji et al. (2014) used vis-NIR spectra and PLSR model in the waterlogged field condition to predict OM, OC, total nitrogen (TN), available phosphorus (AP), available potassium (AK), and pH and obtained good prediction for OM, OC, TN, and pH and poor prediction for AP and AK. In addition, Chang et al. (2005) compared the prediction of total carbon (TC), OC, IC, TN, CEC, pH, soil texture, soil moisture, and potentially mineralizable nitrogen by moist and air-dried samples and indicated an acceptable accuracy of field moist samples. Fystro (2002) obtained higher accuracy of thawed moist samples than dried samples and ground samples when predicting OC and TN. All of these attempts have showed promise in using vis-NIR spectroscopy *in-situ*.

Furthermore, vis-NIR spectroscopy has been used in digital soil mapping (DSM) to rapidly and intensively obtain soil information, simplify the experimental processes of DSM, and improve the accuracy and efficiency of DSM. Brodsky et al. (2013) predicted SOC using vis-NIR spectra and then provided denser input data for DSM. Vasques et al. (2015) used laboratory measured multi-depth spectra for digital soil mapping of soil classes. Rizzo et al. (2016) also used multi-depth vis-NIR spectra to improve the digital soil mapping of soil types. With the development of three-dimensional digital soil mapping (3D-DSM), multi-depth spectral data emerged to play an essential role in improving the reliability of 3D-DSM. Yet, the collection of multi-depth spectral data was mainly implemented on air-dried and ground samples in the laboratory in previous literatures. To our knowledge, only three papers explored the prediction of *in-situ* vis-NIR spectra in soil profiles down to 1 m. Ben-Dor et al. (2008) developed a probe to measure vis-NIR spectra down the profile and used it to predict soil moisture, SOM, carbonates, free iron oxides, and specific surface area (SSA). Viscarra Rossel et al. (2009) assessed the prediction of soil color, mineral composition, and clay content by *in-situ* vis-NIR spectra and reported a good prediction of soil color and mineral composition and even higher accuracy of clay content compared to

laboratory measurement of spectra. Li et al. (2015b) used *in-situ* vis-NIR spectra to predict SOC in soil profiles and resulted in good prediction with an average R^2 of 0.81. Field environment is complex and variable. The deep soil environment is more difficult to control, hence the measurement by inserted spectral probe. Nevertheless, the *in-situ* measurement of multi-depth spectral data holds substantial potential in improving the current techniques and methods of 3D-DSM and requires further exploration.

Therefore, the objectives of this study are 1) to evaluate the prediction of *in-situ* measurement of vis-NIR spectra on a series of soil physical and chemical properties, including volumetric water content (VWC), gravimetric water content (GWC), bulk density (BD), SOM, soil pH, electrical conductivity (EC), sand, silt, clay, available P, and available soil cations (K, Na, Ca, Mg, Zn, Mn, Al, and Fe), and CEC; 2) to assess the performance of the prediction in soil profiles down to 1-m depth.

3.2 Materials and methods

3.2.1 Study area

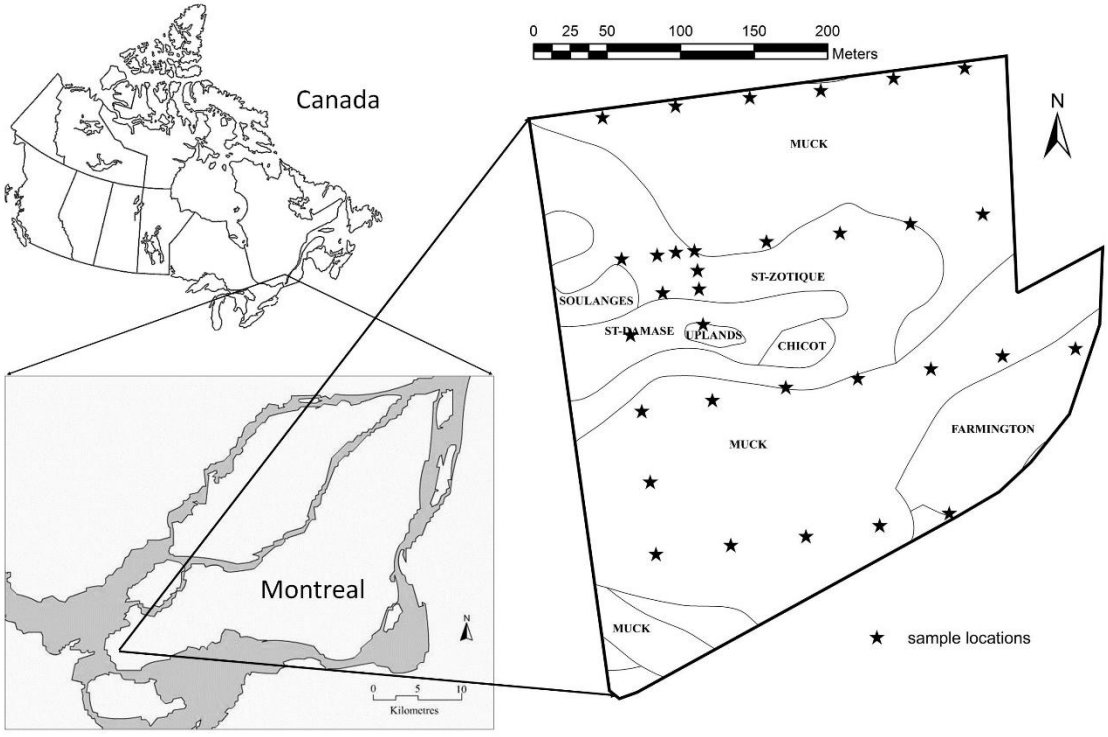

Fig. 3.1. Geographic location of the study area at Macdonald campus of McGill University in Quebec, Canada and locations used to collect soil samples and *in-situ* spectral data.

The study area is located in a small agricultural field (11 ha) of Macdonald Farm, McGill University, Quebec, Canada (45.4 ℃N and 73.9 ℃W) (Fig. 3.1). This field in southern Quebec has experienced various soil forming processes, including glacial deposition, ice retreat, formation of lakes, invasion of saline water, and rises of land level, and resulted in its unique and diverse morphology with various soil types within the field from deep to shallow organic deposits (peat) to mineral soils with highly variable soil textures of sand, sandy loam, silt loam, loam, and clay loam. Soils in this field are classified into multiple soil series including Muck, ST-Zotique, Soulanges, ST-Damase, Uplands, Chicot, Farmington, Chateauguay, Macdonald according to the Canadian Soil Classification System. The elevation ranges from 6.88 to 9.22 meters above sea level. The average annual air temperature over last 30 years is 6.2 ℃ and the average annual precipitation over the last 30 years is 979 mm.

3.2.2 Sample collection and analysis

32 sample locations were identified following a modified nested grid sampling design covering the whole field. *In-situ* near-infrared (vis-NIR) spectra data were collected continuously down to about 1-m depth at these 32 locations using the truck-mounted commercial Veris® P4000 hydraulic soil profiler (Veris Technologies Inc., Salina, KS, USA) in November 2014. At each sample location, 3 (minimum) to 5 vis-NIR spectra profiles (replicates) within 20 cm radius were recorded and averaged to get one set of spectra in order to reduce the interference of instrumental error and small scale spatial variability. Furthermore, the spectra collected within 10-cm depth intervals were averaged at each sampling location for using in the prediction as soil samples were collected and measured at each 10-cm depth intervals. Vis-NIR spectra was measured within a spectral region between 400 nm to 2212 nm at 6-nm intervals, resulting in 371 spectral points (cleaned spectra after removing the edges and the connection in the visible and NIR range).

Soil cores (1-m maximum depth) were collected at these 32 locations using the same hydraulic soil profiler and sectioned at 10-cm depth intervals. A total of 251 samples were collected and sealed in Ziploc bags and transported to the laboratory for analyzing multiple soil properties. All the soil samples were air-dried, ground and sieved to a particle size <2 mm for further analysis. Gravimetric water content (GWC) was determined by the water loss of subsamples which were placed in the oven for 24 hours at 105 ℃. Volumetric water content (VWC) and bulk density (BD) were determined by the water loss of oven-dry samples and the volume of bulk soil samples. Soil organic matter (SOM) was analyzed by loss-on-ignition (LOI) method which calculates the

difference in weight of subsamples after burning at 360 ℃ for 4 hours (Schulte et al., 1991). Soil pH was measured in the soil-water solution to 1:2 soil to water ratio (1:4 for organic soil). Electrical conductivity (EC) was immediately measured after measurement of soil pH by using the same soil-water solution. Particle size distribution (sand, silt, and clay) was determined only for soil samples with SOM < 17% by the hydrometer method (Gee et al., 1986). Soil samples with SOM between 3.5 and 17% were first treated with hydrogen peroxide to remove organic matter, the binding agent. Available P, K, Na, Ca, Mg, Zn, Mn, Al, and Fe were extracted using the Mehlich III solution (a mixture of acetic acid, ammonium nitrate, ammonium fluoride, nitric acid and EDTA) (Ziadi and Tran, 2007). Available P in the extract was determined by a colorimetric technique, and other available cations were determined by an Atomic Absorption Spectrophotometer- Perkin-Elmer 2380. CEC was indirectly calculated by summing all the cations measured by Mehlich III methods.

3.2.3 Spectral preprocessing

All the spectra data was transformed to absorbance (A) before using in the model by calculating the logarithm of the inverse of the reflectance (R) (A=log [1/R]). This is because the absorbance reduces non-linearity in spectra and shows higher correlation with soil properties (Viscarra Rossel et al., 2006c). Soil properties (BD, SOM, GWC, EC, P, K, Na, Ca, Mg, Zn, Mn, and CEC) that were highly skewed were log-transformed (A'=log [A]) to satisfy a normal distribution criteria of the dataset and ensure all the predicted values to be positive. Soil properties (sand, silt, and clay content) whose predicted results were not significantly improved by log-transformation were applied with a normalization method (A"=log[A/(1-A)]) to guarantee the predicted values within the range of 0 to 1 (Diggle et al., 1998). Only four soil properties (VWC, pH, Al, and Fe) were neither log-transformed nor normalized from absorbance. Two different preprocessing approaches were reported to achieve better prediction accuracy, thus tested in this study to determine the best preprocessing approach for specific soil properties, including 1) standard normal variate (SNV) approach that can effectively remove the inferences of varying particle size and light scattering (Barnes et al., 1989) combined with wavelet filter, and 2) Savitzky-Golay filter (Savitzky and Golay, 1964). Additionally, the 1st derivative, which removes the additive constant background effects and enhances spectral feature and mean center, was applied to all the preprocessed spectra (Reeves Iii et al., 2002).

3.2.4 Spectral model (PLSR)

Partial least square regression (PLSR) model was selected to perform the calibration of soil properties against vis-NIR spectra as it is the most commonly used model for spectra analysis in previous literatures. A detailed description and derivation of PLSR can be found in (Geladi and Kowalski, 1986). Basically, PLSR is an improvement of principle component regression (PCR) as it combines the decomposition and regression steps. It reduces the high-collinearity both in the predictors and response variables and uses few factors to explain most of the variations (Viscarra Rossel and Behrens, 2010). In this study, the PLSR models were developed based on the relationship between soil properties and corresponding vis-NIR spectra of total 251 samples. Leave-one-out cross validation (LOOCV) was used to assess PLSR models and select the optimal number of factors.

3.2.5 Validation

Leave-one-out cross validation (LOOCV) was used to select the optimal number of factors (NF) for the spectral model. The accuracy of the LOOCV was determined by root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.1}$$

where *n* was the number of samples, y_i was the measured value, and $\hat{y}_i$ was the predicted value.

In order to avoid over-fitting problems resulting from large factors used in the model, Akaike Information Criterion (AIC) was used to take into account both the accuracy and parsimony of the spectral model (Akaike, 1998) following

$$\mathrm{AIC} = \mathrm{n} \times \log(\mathrm{RMSE}) + 2\mathrm{m} \tag{3.2}$$

where *n* was the number of samples and *m* was the number of PLSR factors in the model. In LOOCV of the spectral model, the NF with the smallest AIC was eventually selected to develop the final calibration model.

Once the NF was determined and the model was fitted. Other criteria, including the coefficient of determination (R^2), mean error (ME), and relative percent deviation (RPD) were calculated to assess the efficiency of the model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.3}$$

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{3.4}$$

$$\text{RPD} = \frac{sd\ of\ y_i}{RMSE} \quad (3.5)$$

where *n* was the number of samples, y_i was the measured value, $\hat{y}_i$ was the predicted value, $\bar{y}$ was the mean of measured values, SS_{res} and SS_{tot} were the sum of squared error of residuals and the total, respectively, sd was the standard deviation of the measured values. R^2 indicated the predictive ability of the spectra model. ME indicated the bias of the predicted values compared to the measured values. Based on the calculated value, RPD was classified into 6 groups representing the performance of the model. RPD<1.0 indicated very poor model prediction; 1.0<RPD<1.4 indicated poor model prediction; 1.4<RPD<1.8 indicated fair model prediction; 1.8<RPD<2.0 indicated good model prediction; 2.0<RPD<2.5 indicated very good model prediction; and RPD>2.5 indicated excellent model prediction (Viscarra Rossel et al., 2006a).

3.2.6 Data analysis

The descriptive statistical characteristics of soil properties were analyzed using Microsoft Office Excel 2013 (Microsoft Inc., Redmond, WA, USA). A nonparametric one-sample Kolmogorov-Smirnov (K-S) test was used to examine the normality of the distribution of soil properties. Basically, it is regarded as a normal distribution when P value of K-S test is over 0.05. The K-S test was conducted in MATLAB R2015b (The Mathworks Inc., MA, USA). The Pearson correlation analysis among soil properties was implemented by the 'corrplot' package (Wei, 2013) in R version 3.2.3 (The R Foundation). The spectra data preprocessing and the PLSR model fitting were conducted using ParLeS version 3.1 (Viscarra Rossel, 2008).

3.3 Results and discussions

3.3.1 Descriptive statistics

All of the soil properties were highly variable within the field with high standard deviations and large ranges (Table 3.1). The majority of the soil properties were positively skewed. This indicated that the observations included many relatively lower values and fewer but scattered higher values especially in GWC with skewness of 4.41, EC with skewness of 5.41, Na with skewness of 4.29, Ca with skewness of 4.28, and CEC with skewness of 3.87. In addition, GWC, EC, Na, Ca, and CEC with kurtosis values of 22.70, 35.40, 20.39, 21.76, and 18.49, respectively further confirmed the extremely skewed distributions. BD, pH, and Fe were slightly negatively skewed. The K-S test indicated that all of the soil properties could not satisfy a normal distribution with *p*-value much smaller than 0.05.

The high variability of soil properties reflected diverse soil pedological features and could be attributed to the influence of various long-term soil forming processes and human interferences. The plow layer of the study field (under agricultural use) was enriched with SOM (mainly in 30%-40%) and contributed to the very dark color of the soil samples. The typical phenomenon that reduced SOM decomposition rate and build-up of SOM in the soil surface resulting from the water-saturated condition of Gleysolic soil (Canadian System of Soil Classification) might have a direct influence in the study area. In addition, Organic soil horizons (Om, Of, and Marl with marine shells inside) were observed at two sample locations that verified the presence of deep organic deposit (peat) and the evidence of the historical presence of Champlain Sea. SOM in these two profiles was as high as 80%. However, the sub and deep soils of most of the soil profiles were dominated by sandy and clayey soils, respectively and exhibited little SOM (mainly lower than 5%) resulting in a dramatic decreasing trend with depth.

Table 3.1 Descriptive statistics of laboratory measured soil properties

Soil Properties	Units	Mean ± Sd[1]	CV[2]	median	min	max	skewness	kurtosis	K-S P value	Normality
VWC	%	44.53±12.56	0.29	44.00	15.00	89.00	0.40	0.44	<0.05	No
GWC	%	**61.67±81.66**	1.32	39.00	9.00	627.00	4.41	22.70	<0.05	No
BD	g cm^{-3}	1.08±0.40	0.37	1.21	0.13	1.74	-0.64	-0.66	<0.05	No
SOM	%	**13.21±19.24**	1.46	1.73	0.30	82.65	1.72	2.20	<0.05	No
pH		7.27±0.66	0.09	7.30	5.36	8.28	-0.32	-0.91	<0.05	No
EC	µS cm^{-1}	**410.51±620.55**	1.51	296.00	47.20	5400.00	5.41	35.40	<0.05	No
Sand	%	32.81±25.62	0.78	32.61	0.05	90.78	0.32	-1.05	<0.05	No
Silt	%	25.84±13.57	0.53	23.09	4.43	57.91	0.54	-0.71	<0.05	No
Clay	%	41.34±25.87	0.63	28.73	3.16	88.89	0.51	-1.18	<0.05	No
P	mg kg^{-1}	**42.40±58.96**	1.39	11.21	0.06	306.74	1.81	3.23	<0.05	No
K	mg kg^{-1}	129.93±99.59	0.77	78.94	19.07	412.40	0.86	-0.63	<0.05	No
Na	mg kg^{-1}	**85.24±102.33**	1.20	63.28	15.21	764.90	4.29	20.39	<0.05	No
Ca	mg kg^{-1}	**1955.78±4419.13**	2.26	496.37	91.77	32770.82	4.28	21.76	<0.05	No
Mg	mg kg^{-1}	650.84±363.85	0.56	583.46	74.96	1803.36	1.00	0.81	<0.05	No
Zn	mg kg^{-1}	**7.26±9.17**	1.26	3.10	0.43	45.82	2.05	3.90	<0.05	No
Mn	mg kg^{-1}	**20.06±21.36**	1.06	8.26	0.98	89.26	1.27	0.42	<0.05	No
Al	mg kg^{-1}	620.93±236.20	0.38	631.51	15.04	1453.81	0.05	0.39	<0.05	No
Fe	mg kg^{-1}	317.92±83.90	0.26	326.00	94.34	681.64	-0.07	1.11	<0.05	No
CEC	meq $100g^{-1}$	24.54±22.45	0.92	19.41	6.79	175.75	3.87	18.49	<0.05	No

[1]Sd: standard deviation; [2]CV: Coefficient of variation; the bold indicates the standard deviations are larger than the mean values.

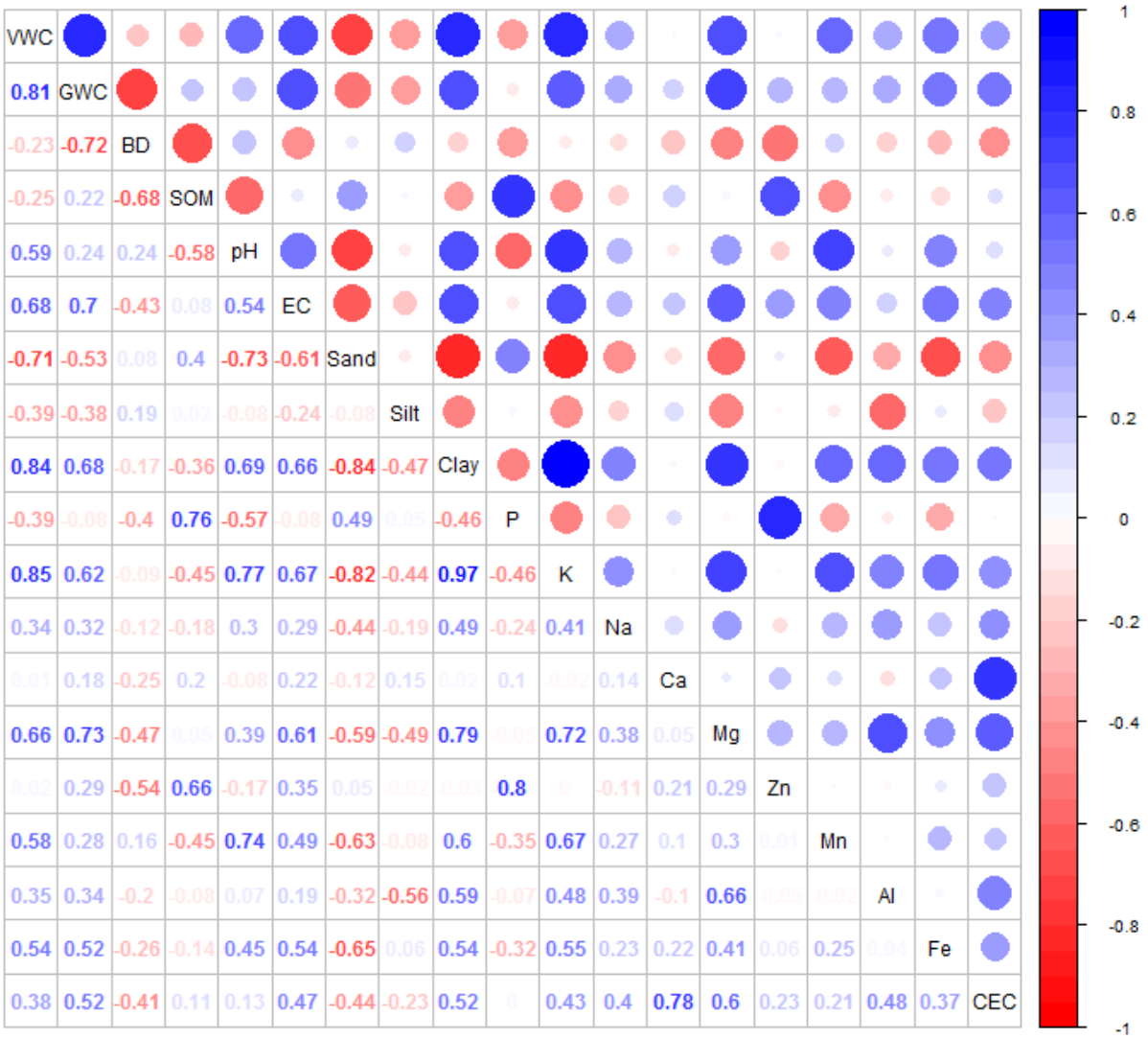

Fig. 3.2. Pearson correlation coefficients amongst laboratory measured soil properties.

The correlation relationship amongst soil properties was shown in Fig. 3.2. Several soil properties were highly correlated with SOM. SOM is essential in forming soil aggregates, increasing porosity, keeping good soil structure, maintaining soil water, and absorbing soil nutrients, thus showing a negative correlation with BD and positive correlation with P and Zn. Additionally, the decomposition of SOM released organic acids and reduced soil pH. However, the weak correlations between SOM and VWC and GWC may be due to the interferences by clay content. Clay with small pores in the deep soil can hold a large amount of water with positive correlation coefficients of 0.86 and 0.74 with VWC and GWC, respectively but with very little presence of SOM in the clay soil. The vast clay plain presented in deep soil was another evidence of the Champlain Sea and it was formed during the marine deposition (Chapman and Putnam, 1984). In

addition, clay was positively correlated with pH (0.65), EC (0.73), K (0.96), Na (0.6), Mg (0.81), Mn (0.68), Al (0.59), Fe (0.54), and CEC (0.52) may be due to the absorbing capacity and attraction to base cations. Positive correlations were also observed amongst these properties to some extent. However, sand content was highly negatively correlated (-0.86) with clay content and thus negatively correlated with VWC (-0.79), GWC (-0.65), and other properties associated with base cations. The negative correlation might also be due to the presence of large pores and smaller specific surface area of sand. The high variability of these soil properties led to challenges, while providing opportunities to test the predictive ability of the spectral models of *in-situ* measurements. Furthermore, the relationship among the soil properties played an essential role in interpreting the predictive performance of spectral models.

3.3.2 Description of vis-NIR spectra

The main processes in visible region were electronic transitions, while the NIR spectra was dominated by weak overtones and combinations of fundamental vibrations of molecular bonds in the mid-infrared (MIR) region (Stenberg et al., 2010). The absorption features were determined by specific soil constituents or molecular functional groups and the qualitative features of the spectra were analyzed by positive or negative peaks at specific wavelengths in the characteristic shape (Miller, 2001). Basically, all the soils showed similar spectra with fewer absorption features but the absorbance of spectra holistically increased with the increase of SOM content (Fig. 3.3). The broad and well recognized absorption features in the vis-NIR spectra were at around 1450 nm and 1920 nm indicating the presence of O-H bond from water molecules (Ramirez-Lopez et al., 2014). Important absorption features of organic soils but different from mineral soils were the absorptions at 415 nm, 570 nm, and 660 nm representing the characteristics of SOC (Shonk et al., 1991). The clear peaks at 1000 nm and 1170 nm could be attributed to the overtones of N-H and C-H bonds which contributed to the prediction of organics (Viscarra Rossel and Behrens, 2010). The removal of bands from 1018 nm to 1075 nm resulted in abrupt changes in Fig. 3.3. In addition, the broad absorption bands in the visible region, determined by chromophores and darkness of SOM, was another typical characteristic of the organic soil (Stenberg et al., 2010). The positive peaks at around 460 nm, 540 nm, and 650 nm allowed for the detection of sand content (Viscarra Rossel et al., 2006c). Gholizade et al. (2013) observed strong positive correlation of EC and P with absorption bands at 440 nm and 490 nm and pH with absorption band at 850 nm.

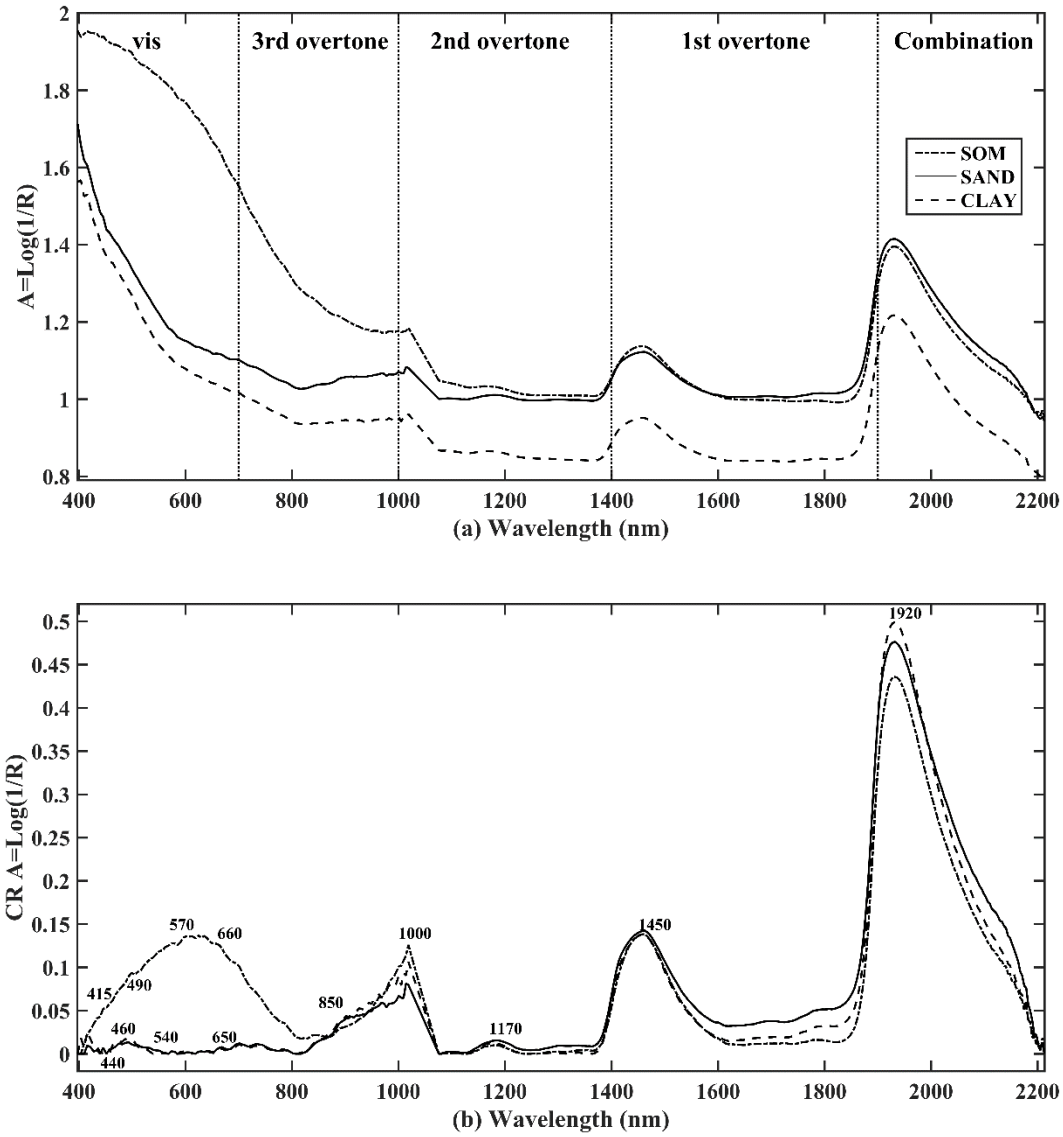

Fig. 3.3. (a) Examples of collected vis-NIR spectra of three typical soil types, 1) organic soil collected in the top soil horizon with SOM content of 38.35%; 2) sandy loam soil collected in the subsoil horizon with SOM content of 1.07% and sand content of 64.16%; 3) clay soil collected in the deep soil horizon with SOM content of 0.96% and clay content of 78.36%. (b) Corresponding continuum-removed (CR) absorbance spectra.

3.3.3 PLSR model

Table 3.2 Results of spectral models for multiple soil properties

Soil Properties	Pretreatment	NF	R^2	ME	RPD
VWC	SNV[1]	7	0.667	-0.000	1.73
GWC	Log[2]+SNV	8	0.740	-0.000	1.96
BD	Log+SG[3]	6	0.760	-0.000	2.04
SOM	Log+SG	13	0.775	-0.002	2.10
pH	SNV	13	0.657	-0.002	1.70
EC	Log+SNV	8	0.750	0.000	2.00
Sand	Normalization[4]+SNV	8	0.538	0.009	1.47
Silt	Normalization+SNV	5	0.187	0.002	1.11
Clay	Normalization+SNV	8	0.594	-0.002	1.57
P	Log+SNV	12	0.711	0.003	1.86
K	Log+SG	8	0.528	0.002	1.45
Na	Log+SG	7	0.587	0.002	1.56
Ca	Log+SNV	9	0.650	-0.001	1.69
Mg	Log+SNV	9	0.586	0.002	1.55
Zn	Log+SNV	4	0.553	0.003	1.50
Mn	Log+SNV	7	0.553	-0.005	1.50
Al	SNV	11	0.451	-2.170	1.34
Fe	SG	11	0.388	0.231	1.27
CEC	Log+SNV	9	0.649	-0.000	1.69

[1]SNV: standard normal variate (SNV) approach with wavelet filter; [2]Log: log-transformation to achieve a normal distribution (A'=log[A]); [3]SG: Savitzky-Golay filter; [4]Normalization: normalization method to guarantee the predicted values within the range of 0 to 1(A"=log[A/(1-A)]).

Due to the skewness of the distribution of most of the soil properties (Table 3.1), log-transformation and normalization processes were implemented before the model fitting. SOM, as the most frequently predicted property by vis-NIR spectra, was also well predicted with R^2 of 0.78 and RPD of 2.10 in this study, a slightly worse than others with R^2 above 0.8 (Chang and Laird, 2002; Chang et al., 2001; Fidencio et al., 2002). This may be attributed to the large range (0.30%-82.65%) of the SOM content and diverse soil textures in this study. Another study with large SOM variation e.g. 0.01%-59.34% from Canadian prairie region (Malley et al., 2000) also reported similar R^2 of 0.78 and RMSE of 29. The weak absorption features in NIR regions were mainly at 1100, 1600, and 1700-1800 nm resulting from the overtones and combinations of NH, CH, and CO groups (Ben-Dor et al., 1999; Clark, 1999). However, these features were not sufficient and

efficient for identifying and predicting SOM due to the over-lapping bands in the NIR region. The dark color of SOM can be clearly detected by the broad absorption feature in the visible region that greatly improved the prediction of SOM by vis-NIR spectra (Viscarra Rossel et al., 2006b). Owing to the very clear absorption features of O-H water bond, VWC with R^2 of 0.67 and RPD of 1.73 and GWC with R^2 of 0.74 and RPD of 1.96 were relatively well predicted. In addition, the soil water related properties also showed relatively better prediction, including BD with R^2 of 0.76 and RPD of 2.04 which showed negative correlation with GWC (-0.72) and EC with R^2 of 0.75 and RPD of 2.00 which showed positive correlation with VWC (0.68) and GWC (0.7). The BD was also influenced by SOM due to the negative correlation (-0.69) and EC was also influenced by clay content due to the positive correlation (0.73).

Clay, as a fundamental constituent of soil, played an essential role in maintaining soil structure and forming soil aggregates (Stenberg et al., 2010). Clay content with the large range from 3.2% to 88.9% was fairly well predicted with R^2 of 0.59 and RPD of 1.57. The absorption features of clay content were usually very strong and mainly determined by O-H in water molecules and Mg-, Al-, and Fe-OH in the minerals (mainly displayed in the 2200-2500 nm region) (Ben-Dor and Banin, 1995). However, in this study, the vis-NIR region only reached 2200 nm, so that the prediction was not as good as other studies with R^2 more than 0.60 and even as high as 0.94 (Stenberg et al., 2002). The sand content with the range of 0.05%-90.78% was fairly predicted with R^2 of 0.54 and RPD of 1.47, which was not significantly different from other studies with R^2 mainly within 0.5-0.6 (Brown et al., 2006; Islam et al., 2003). The optical characteristics of sand content is mainly determined by the amount of quartz and could contribute to the weaker prediction (White et al., 1997). The silt content with relatively narrow range of 4.43%-57.91% was poorly predicted with R^2 of 0.19 and RPD of 1.11 compared to previous studies with R^2>0.8 (Chang et al., 2001; Vendrame et al., 2012). While, extremely poor prediction of silt with R^2 of 0.05 and RPD of 0.9 was also observed by Islam et al. (2003). Silt doesn't have strongly recognizable feature in vis-NIR region, and its correlation with other fundamental soil properties was really weak in this study, thus resulting in poor prediction.

Wu et al. (2010) classified soil properties into primary, secondary, and tertiary properties according to their responses to vis-NIR spectroscopy. SOM and clay content were primary and secondary properties which are essential composition of soil and show clear absorption features in characteristic curve. While tertiary properties are usually predicted based on their correlation with

primary and secondary properties. Soil pH, available P, K, Na, Ca, Mg, Zn, Mn, Al, Fe, and CEC were not expected to have good prediction attributed to their indirect relationship with vis-NIR spectra. However, on the basis of their positive or negative correlation with other soil constituents (mainly SOM and clay), diverse prediction results were achieved with R^2 ranging from very low to very high (Stenberg et al., 2010). Soil pH was predicted with R^2 of 0.66 and RPD of 1.70 due to its negative correlation with SOM and sand and positive correlation with EC and clay. Available P, an important but usually limited plant nutrient, has been widely studied for its field variability and predictive ability from various sensors for precision agriculture. Prediction results of P from vis-NIR were usually highly variable in previous studies. In this study, a good model prediction with R^2 of 0.71 and RPD of 1.86 was achieved. P was positively correlated with SOM (0.65) and the good prediction of SOM might have contributed to the good result of prediction for P. Daniel et al. (2003) demonstrated that elements e.g. P and K are not optically active to stimulate reflectance variation while they can be indirectly predicted based on the correlation with other soil properties. In addition, the Mehlich III extracted P was more suitable for spectral model calibration compared to other extraction methods (Chang et al., 2001). Ca was observed with good prediction with R^2 of 0.65 and RPD of 1.69, while slight worse than the study with R^2 of 0.80 and RPD of 2.19 (Chang et al., 2001). However, the range of Ca in this study is 91.77-32770.82 mg kg^{-1} much larger than that in the previous study with a range of 87.7-12763.9 mg kg^{-1}. Ca is only highly correlated with CEC (0.78) and a comparable good prediction of CEC with R^2 of 0.59 and RPD of 1.69 was observed. These prediction results of Ca and CEC are similar to the results obtained by Islam et al. (2003) with RPD of 1.7 and 1.6, respectively. Moderate prediction results of soil available K, Na, Mg, Zn, and Mn were obtained in this study with R^2 values of 0.53, 0.59, 0.59, 0.55, and 0.55, and RPD values of 1.45, 1.56, 1.55, 1.50, and 1.50, respectively, mainly owing to their positive relationship with VWC, GWC, and clay content. However, poor results were obtained for Al and Fe with R^2 of 0.45 and 0.39 and RPD of 1.34 and 1.27, respectively. Terra et al. (2015) observed similar performance with R^2 of 0.43 and 0.39 for Al and Fe, respectively, while a better result of Fe with R^2 of 0.64 was observed by Chang et al. (2001). Generally, prediction of cations was usually uncertain and largely determined by their correlation with other soil properties.

3.3.4 Depth specific prediction performance

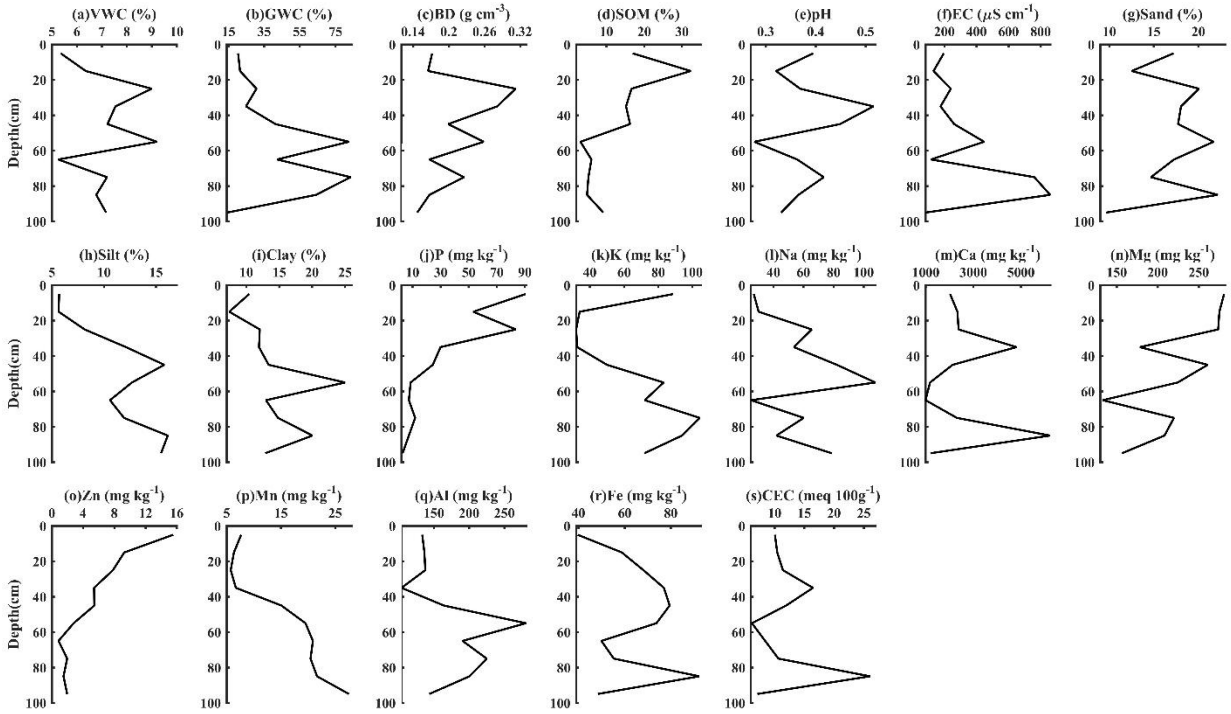

Fig. 3.4. Depth specific prediction performance (RMSE) of spectral models for diverse soil properties.

The depth specific prediction performance was shown as the changes of RMSE values with depth in Fig. 3.4. The number of samples vary for different depths which might have influence on the performance. Although most soil properties did not show specific tendency of prediction accuracy with depth, a decreasing trend was clearly observed for SOM, P, and Zn and an increasing trend was also obvious for silt content and Mn. The decreasing trend of RMSE values was mainly determined by the drastic decline of SOM, P, and Zn with depth and the relatively narrow range of soil properties in deep soil. Li et al. (2015b) observed a similar decline of correlation of SOC and vis-NIR spectra and RMSE values in leave-one-out cross validation results with depth and they attributed these phenomena to the decrease of SOC with depth. Similarly, the increasing trend of soil properties was due to the increasing of silt content and Mn with depth. However, other soil properties displayed unpredictable changes with depth. Despite the complex environment and measuring difficulty in deep soil, it did not certainly lead to the worst results, and even slightly better than sub soil for VWC, BD, pH, and Mg. In this field, the top soil is uniform due to the mixing of tillage and deep soil is also uniform due to the less disturbance and fluctuating water table, while the sub soil is relatively complex and changeable. This might be partly responsible for

the worse results in sub soil of VWC, BD, pH, and Mg. Some properties (clay content, K, Ca, Al, Fe, and CEC) displayed worse results in deep soil compared to top and sub soil. This could be attributed to the high clay content in deep soil and other properties which showed highly positively correlation with clay content were also plenty in deep soil. Therefore, the depth did not show directly influence the prediction of spectral model, whereas it determined the soil properties (magnitude and range) and hence the prediction accuracy. Accordingly, RMSE values were largely affected by the actual values of soil properties.

3.4 Conclusion

The results indicated that *in-situ* measurement of vis-NIR spectra can be used to predict diverse soil properties with good accuracy. Soil vis-NIR spectra with strong and easily recognizable absorption features of SOM and water, could deliver good predictions for SOM, GWC, VWC, as well as their highly correlated soil properties (BD, EC, and P). Due to the missing of essential spectral bands between 2200 and 2500 nm, the prediction of clay content was not as good as previous studies. Other soil properties (sand content, pH, K, Na, Ca, Mg, Zn, Mn, and CEC), with direct or indirect relationships with SOM and soil water, were reported with fair prediction results with R^2 above 0.5. However, silt content, Al, and Fe were observed with relatively poor prediction. In short, soil vis-NIR spectra is a good technique to easily and exhaustively obtain soil property information *in-situ*. The depth specific prediction performance indicated that depth did not directly exaggerate the operational error and reduce prediction accuracy of spectral model, but it influenced on the magnitude and range of soil properties which greatly affected the accuracy.

PREFACE TO CHAPTER 4

Based on the developed vis-NIR spectral models in Chapter 3, 14 soil properties (including VWC, GWC, BD, SOM, soil pH, EC, sand content, clay content, available P, K, Na, Mg, Zn, and Mn) with good prediction were finally chosen for the actual 3D-DSM in Chapter 4. A thorough understanding of all these soil properties helped to uncover the associated relationships among them and interpret the distribution of final maps. Exhaustive soil information (148 soil profiles) were obtained by vis-NIR spectral models developed in Chapter 3 which were further separated to calibration dataset for developing 3D-DSM and validation dataset for testing the accuracy of models. Three regression techniques discussed in Chapter 2 were assessed and compared for the mapping accuracy and the best one was chosen for the final maps.

Chapter 4 has been written as a research paper format and will be submitted to *Geoderma* (impact factor 2.85). The detailed information for authors is shown below:

Order of authors: Yakun Zhang[a], Asim Biswas[a,*], Wenjun Ji[b], & Viacheslav I. Adamchuk[b].

The author of this thesis took charge of experimental design, field experiment, laboratory measurement, data analysis and interpretation, and manuscript preparation. Prof. Biswas as thesis supervisor was completely involved in this study and greatly assisted in the experimental design, technical support, scientific advice, and review and edition of the manuscript. Wenjun helped with the vis-NIR spectra data preparation and advised on the spectral model development. Prof. Adamchuk as thesis committee provided much support on the technical facilities, environmental variables collection and compilation, and data analysis.

CHAPTER 4

Three-dimensional digital soil mapping of multiple soil properties at a field-scale using 3D regression kriging

Abstract:

With technological advancements and development of computational facilities, 3D digital soil mapping (DSM) is becoming popular for its information on both the horizontal and the vertical variability in soil properties. Most current studies are based on either, one-dimensional profile depth functions, or two-dimensional horizontal interpolation techniques, which do not allow true 3D visualization of spatial soil heterogeneity. Only few studies have utilized the 3D variograms for mapping. Recent advances in proximal soil sensing technologies allow measurement and prediction of soil properties rapidly at multiple depths which could serve as input for DSM. Various soil physical and chemical properties have already shown either direct or indirect relationships with the proximal soil sensing data, including volumetric water content (VWC), gravimetric water content (GWC), bulk density (BD), soil organic matter (SOM), soil pH, electrical conductivity (EC), sand content, clay content, available phosphorus (P), and available soil cations (potassium (K), sodium (Na), magnesium (Mg), zinc (Zn), and manganese (Mn)). This study aims to test the methodology of 3D-DSM using a 3D geostatistical approach with predicted soil properties from proximal soil sensing. In this study, 32 soil cores (1-m maximum depth) were collected and sectioned at 10-cm depth intervals from Field 26 of Macdonald Farm, McGill University. A total of 251 samples were analyzed for multiple soil properties in the laboratory at McGill University. Additionally, vis-NIR spectra data were collected to about 1-m depth *in-situ* at 148 sites (including these 32 cores) using the Veris® P4000 soil profiler. Predicted soil properties by partial least square regression (PLSR) models from vis-NIR spectra at 148 sites were separated into calibration dataset (70%) and validation dataset (30%). The spatial relationship of soil properties from calibration dataset and environmental covariates (including field topography, gamma-ray radiation, and apparent soil electrical conductivity) was used to estimate 3D variograms and pursue regression kriging. Generalized linear model, regression tree, and random forest were compared for the trend prediction. As a result, complete three-dimensional digital soil maps were developed.

4.1 Introduction

With dramatically increasing global issues, such as overpopulation, food security, climate change, environmental pollution and degradation, and natural resource depletion, digital soil mapping (DSM) has become a powerful approach in assisting optimal decisions on environmental and agricultural management by providing available soil information (Arrouays et al., 2014). The DSM aims to predict continuous, quantitative soil properties and visually display the spatial variability by incorporating high-resolution digital soil sensing and mapping techniques (McBratney et al., 2003).

Information on variability in deep soil is essential for interpretation and management of soil drainage, stable soil carbon storage, leaching dynamics, and groundwater associated studies but has been rarely studied due to technical constraints. During the last decade, with the development of computational methods and technical advances, three-dimensional digital soil mapping (3D-DSM) has been popularly explored for its great potential in determining both the horizontal and the vertical variability in soil properties. Most current studies created 3D maps by integrating one-dimensional profile depth functions and two-dimensional horizontal regression or interpolation techniques and was denoted as a 'top down' method (Veronesi et al., 2012). For example, Veronesi et al. (2012) combined polynomial function and ordinary kriging (OK) for mapping soil compaction in 3D. Liu et al. (2013) combined equal-area quadratic spline function (EAQSF) and radial basis function (RBF) neural network for 3D mapping of soil organic carbon (SOC). Lacoste et al. (2014) applied EAQSF and cubist method for 3D SOC mapping. Kidd et al. (2015b) and Viscarra Rossel et al. (2015) combined EAQSF with cubist model and residual kriging for 3D mapping of multiple soil properties in Australia. Taghizadeh-Mehrjardi (2016) utilized EAQSF and genetic programming (GP) for 3D cation exchange capacity (CEC) mapping. Liu et al. (2016) examined the application of linear function and power function as a profile depth function combined with interpolation technique for 3D soil organic matter (SOM) mapping. Taghizadeh-Mehrjardi et al. (2016) compared six 2D data mining methods with EAQSF for 3D mapping of SOM. However, this 'top down' method quantifies vertical and horizontal variability separately and could not completely uncover the 3D interactions thus missing the true representation of spatial soil heterogeneity in 3D. In addition, though the flexibility of EAQSF allows to mathematically fit the vertical variability of multiple soil properties with high accuracy, it fails to consider the real soil physical conditions and to understand soil profile morphology. Whereas more

specific profile depth functions e.g. exponential decay function for SOM limits its use for other soil properties and reduces the generality.

Another commonly used 3D mapping technology is the direct use of 3D geostatistical interpolation methods. A 3D variogram was fitted in 3D universal kriging (UK) for mapping soil textures and it achieved the most accurate results compared to the 'top down' methods (Veronesi, 2012). Poggio and Gimona (2014) developed 3D mapping of SOC stocks by combining a 3D model for trend prediction and a 3D kriging. Brus et al. (2016) used 3D geostatistical modelling for SOC by combining a linear mixed model with covariance analysis of regression residuals with the consideration of their interactions with depth. Moreover, an automated mapping procedure using 3D regression kriging (RK) was developed for producing global maps of various soil properties (Hengl et al., 2014). This greatly optimized the methodology of 3D variogram by involving depth functions and soil anisotropy relationships thereby simplifying the whole mapping and updating procedure. Additionally, 3D RK provides an opportunity to test the linear or nonlinear relationships of soil properties against environmental covariates by various data mining methods and increase the prediction accuracy but requires further application and detection.

Emerging proximal soil sensing platforms can rapidly obtain soil data at multiple depths and, therefore, be used to predict many soil properties at depths. Diffuse reflectance spectroscopy e.g. vis-NIR spectra (400-2500 nm) has gained attention of soil scientists in recent past as a good alternative to conventional laboratory based soil analysis for its advantages such as rapid, simple, non-destructive, accurate, and cheap measurement of soil properties (Stenberg et al., 2010). For example, Brodsky et al. (2013) have utilized vis-NIR spectra to predict SOC which was further used as an input in DSM. This substantially reduced the workload of the integral mapping process and increased the resolution and accuracy by providing relatively denser input data. Rizzo et al. (2016) used multi-depth vis-NIR spectral library to map soil color and soil type. Soil properties including SOM and clay content are fundamental soil constitutes and have clear absorption features in the vis-NIR region that have been successfully used to predict those properties (Rossel and Lark, 2009). Soil moisture, with strong and easily recognized absorption feature due to the O-H functional group, has been predicted with greater confidence (Stenberg et al., 2010). Other soil properties, including soil pH, plant nutrients, and CEC, which are dependent on SOM, soil texture, and other soil attributes, thus indirectly determined by vis-NIR spectra, have shown diverse and unstable prediction accuracy but with potential of future exploration (Shepherd and Walsh, 2002;

Vendrame et al., 2012). In addition, even though the *in-situ* measurement of spectra data which was influenced by the field moisture and weather condition was reported a poorer predictive ability compared to lab-based spectra data (Daniel et al., 2003), it is a rapid and real-time measurement approach with tremendous potentiality and has attracted much attention recently.

The objectives of this study were: 1) to use exhaustively predicted soil properties from *in-situ* vis-NIR spectra as input for DSM; 2) to produce the 3D digital soil maps of multiple soil properties using the state-of-the-art 3D RK method; 3) to assess the prediction ability of three different linear and nonlinear regression methods, including generalized linear model (GLM), regression tree (RT), and random forest (RF) for the tendency prediction of 3D RK.

4.2 Materials and methods

4.2.1 Study area, sample collection and processing

This study was conducted in a small farm (11 ha), McGill University, Quebec, Canada (45.4 °N and 73.9 °W) (Fig. 4.1). Soil in this field is highly variable with soil types ranging from organic deposits (peat) to mineral soils (sandy, sandy loamy, silt loamy, loamy, and clayey soil). 148 sample locations were identified by incorporating six sampling designs including grid sampling, grid random sampling, simple random sampling, stratified random sampling, transect sampling, and conditioned Latin hypercube sampling. Out of these, 32 locations were selected following a modified nested grid sampling design. Near-infrared (vis-NIR) spectra data were collected continuously down to about 1-m depth at these 148 locations in the field condition using the truck-mounted commercial Veris® P4000 hydraulic soil profiler (Veris Technologies Inc., Salina, KS, USA) in November 2014. Additionally, soil cores were collected to about 1-m depth at the 32 locations by the same soil profiler and sectioned at 10-cm depth intervals, resulting in 251 soil samples. A detailed description of *in-situ* spectral data collection and processing, laboratory analysis of soil samples, and the development of spectral models (partial least square regression (PLSR)) of soil properties and vis-NIR spectra at 32 locations could be found in chapter 3.

VWC, GWC, BD, SOM, soil pH, EC, sand content, clay content, P, K, Na, Mg, Zn, and Mn have been proven to be accurately predicted from vis-NIR spectra with $R^2 > 0.50$. Therefore, the developed spectral models of were further used to predict these 14 soil properties at total 148 locations and prepare for DSM. Afterwards, 45 sample locations (30%) were selected based on probability sampling and reserved for validation, and the remaining 103 sample locations (70%) were used for building DSM.

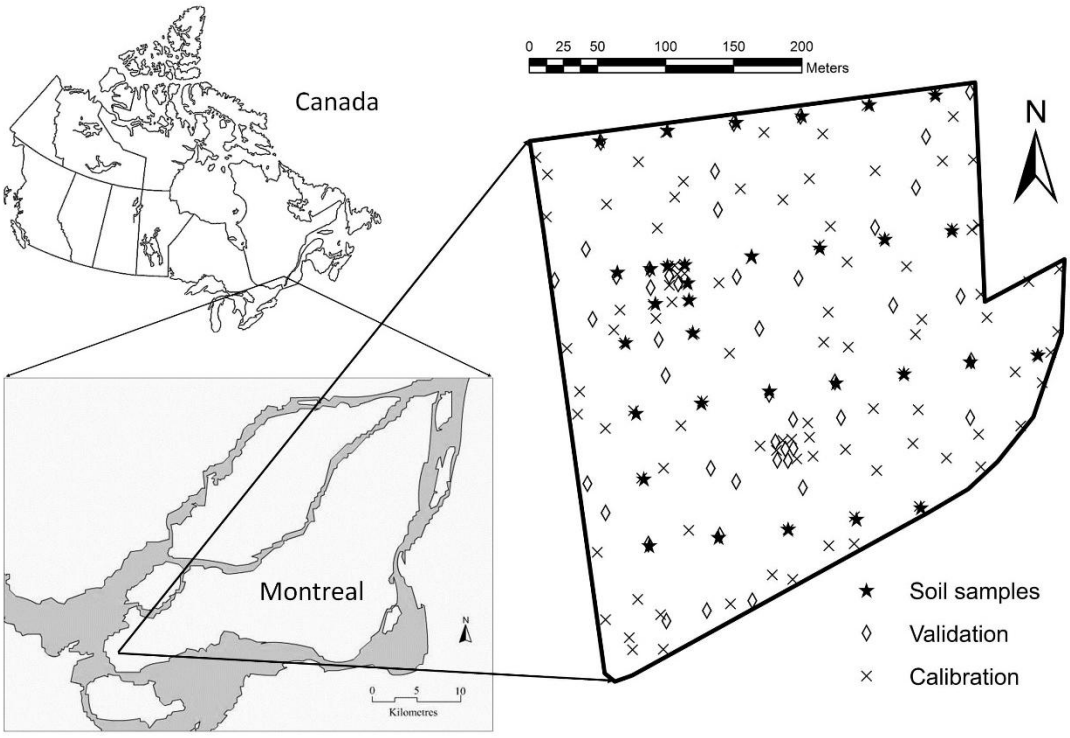

Fig. 4.1. Study area of Macdonald Farm of McGill University in Quebec, Canada, locations used to collect soil samples and in-situ spectral data, and division of calibration and validation dataset.

4.2.2 Environmental covariates

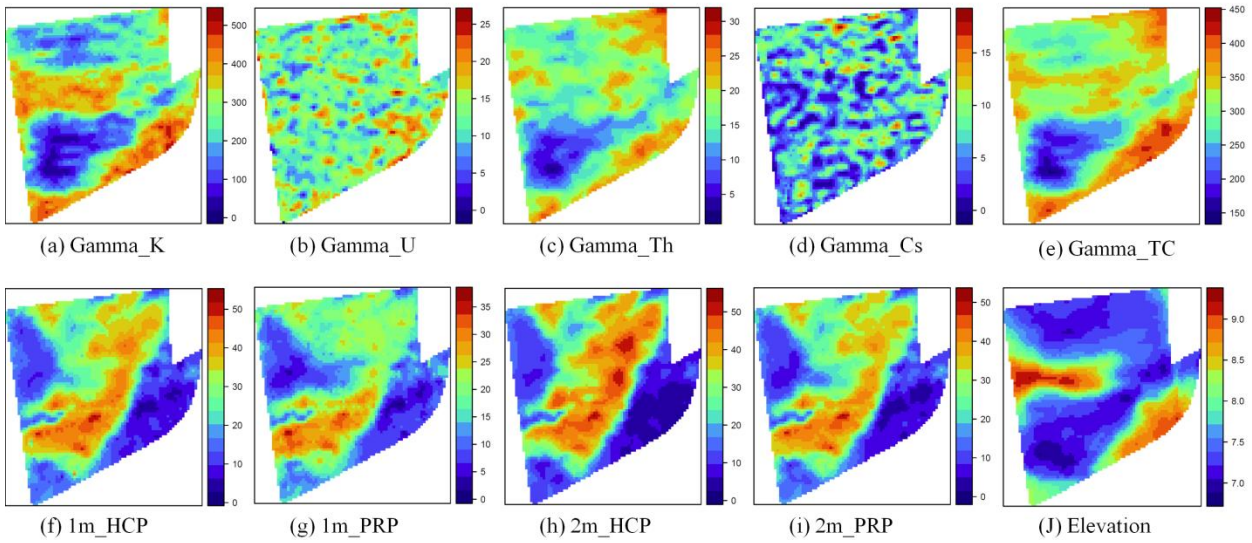

Fig. 4.2. Environmental covariates used in DSM. Gamma_K indicated Potassium-40. Gamma_U indicated Uranium-232. Gamma_Th indicated Thorium-238. Gamma_Cs indicated Caesium. Gamma_TC indicated total radiometric count. These were measured by gamma-ray spectrometer.

1m_HCP indicated horizontal coplanar at 1 m distance of DUALEM-21S. 1m_PRP indicated perpendicular coplanar at 1.1 m distance of DUALEM-21S. 2m_HCP indicated horizontal coplanar at 2 m distance of DUALEM-21S. 2m_PRP indicated perpendicular coplanar at 2.1 m distance of DUALEM-21S. Elevation was measured by Real Time Kinematic (RTK).

On the basis of the soil-landscape scorpan model (McBratney et al., 2003):

$$S_a \; or \; S_c = f(s, c, o, r, p, a, n) \tag{4.1}$$

where S_a represented soil attributes maps, S_c represented soil class maps, s indicated soil (soil attributes, remote and proximal soil sensing), c indicated climate (temperature, precipitation), o indicated organisms (microbial activities, vegetation), r indicated relief (topography, landscape attributes), p indicated parent material (lithology), a indicated age (time), and n indicated space (geographical coordinates). Ten environmental covariates (Fig. 4.2) mainly representing s and r attributes were selected for DSM in this study. Five covariates derived from gamma-ray spectrometer includes Potassium (^{40}K), Uranium (^{232}U), Thorium (^{238}Th), Caesium (^{137}Cs), and total radiometric count (TC). It measured the decay of natural radioelements at top soil (0-30 cm). The gamma-ray spectrometer was installed on a truck, 0.5 m above the ground. Measurements were continuously collected along lines with row width of 11 m at an average travel speed of about 6.2 km/h. Four covariates were derived from apparent soil electrical conductivity measurements by Electromagnetic Induction (EMI) instrument DUALEM-21S (Dualem, Inc., Milton, ON, Canada). The DUALEM-21S consisted of a 2.41-m long tube, a transmitter coil, and four receiving coils. The receiving coils provided information on horizontal coplanar at 1 m (1m_HCP) and 2 m (2m_HCP) distances from transmitter and perpendicular coplanar at 1.1 m (1m_PRP) and 2.1 m (2m_PRP) distances from transmitter. In addition, the effective sensing depth (75% response) of 1m_HCP, 2m_HCP, 1m_PRP, and 2m_PRP were 1.55 m, 3.18 m, 0.54 m, and 1.03 m, respectively. The sampling of DUALEM-21S was implemented along lines with row width of 12 m at an average travel speed of 5 km/h. Last environmental variable is elevation collected by Real Time Kinematic (RTK) GNSS receiver (Trimble RTK/PP-4700 GPS, Trimble Navigation Limited, Sunnyvale, CA, USA). However, only eight of the ten covariates were used in the mapping after excluding Gamma-U and Gamma-Cs due to the poor spatial structure (Fig. 4.2). All the covariates were exhaustively collected at point locations in October 2013. The measurements of ten environmental covariates were individually interpolated into raster maps by ordinary kriging and resampled at 5-m resolution in ArcGIS 10.3.1 (ESRI Inc.). A principle component analysis

(PCA) was implemented on the environmental covariates to reduce the collinearity of predictors and obtain independent components (Hengl et al., 2003) prior to the DSM in R version 3.2.3. Eight principle components (PCs) derived from PCA as well as altitudes were subsequently used in the DSM.

4.2.3 Digital soil mapping (DSM)

Regression kriging (RK) is an approach that integrates regression technique and interpolation method in a single step so that both the relationship between target soil properties and explanatory variables and spatial structure of the variables could be simultaneously interpreted (Hengl et al., 2007). The RK can be expressed as

$$\hat{z}(\boldsymbol{s}_0) = \sum_{j=0}^{p} \hat{\beta}_j \cdot X_j(\boldsymbol{s}_0) + \sum_{i=1}^{n} \lambda_i \cdot e(\boldsymbol{s}_i) \quad (4.2)$$

where $\sum_{j=0}^{p} \hat{\beta}_j \cdot X_j(\boldsymbol{s}_0)$ represents the trend prediction by regression and $\sum_{i=1}^{n} \lambda_i \cdot e(\boldsymbol{s}_i)$ represents the regression residual prediction by kriging. More specifically, $\hat{z}(\boldsymbol{s}_0)$ is the predicted value of the target soil property at location $\boldsymbol{s}_0$; $X_j(\boldsymbol{s}_0)$ is the environmental covariate in location $\boldsymbol{s}_0$; $\hat{\beta}_j$ is the regression coefficients of the regression of $\hat{z}(\boldsymbol{s}_0)$ on $X_j(\boldsymbol{s}_0)$; p is the number of predictors; λ_i is the kriging weight of residual; $e(\boldsymbol{s}_i)$ is the regression residual at location $\boldsymbol{s}_i$; and n is the number of observations.

This two-dimensional (2D) RK was extended to 3D RK by incorporating the depth function and depth parameters (Hengl et al., 2014) and was written as

$$\hat{z}(\boldsymbol{s}_0, d_0) = \sum_{j=0}^{p} \hat{\beta}_j \cdot X_j(\boldsymbol{s}_0, d_0) + \hat{\boldsymbol{g}}(d_0) + \sum_{i=1}^{n} \lambda_i(\boldsymbol{s}_0, d_0) \cdot e(\boldsymbol{s}_i, d_i) \quad (4.3)$$

where $\left(\sum_{j=0}^{p} \hat{\beta}_j \cdot X_j(\boldsymbol{s}_0, d_0) + \hat{\boldsymbol{g}}(d_0) \right)$ indicated trend prediction of the regression and $(\sum_{i=1}^{n} \lambda_i(\boldsymbol{s}_0, d_0) \cdot e(\boldsymbol{s}_i, d_i))$ indicated the kriging of residuals. Unlike 2D RK, all the locations $\boldsymbol{s}_0$ were extended to $(\boldsymbol{s}_0, d_0)$ by combining both geographical coordinates $\boldsymbol{s}_0$ and depth parameters d_0. $\hat{\boldsymbol{g}}(d_0)$ was the depth function and $X_j(\boldsymbol{s}_0, d_0)$ was the value of predictors (PCs derived from environmental covariates in this study) at location $(\boldsymbol{s}_0, d_0)$. An equal-area quadratic spline function was chosen as depth function in this paper. In addition, a new profile depth function was developed in appendix A to specifically quantify the vertical variability of soil pH. Although the new model was not applied in this paper, it would be further tested for its feasibility in 3D RK.

Various linear and nonlinear regression techniques were applied to assist the tendency prediction. In this study, generalized linear model (GLM), regression tree (RT), and random forest (RF) were compared and assessed for their prediction ability. GLM modeled the linear relationship between

responses and predictors by ordinary least square fitting and allowed various residual distributions other than a normal distribution (Venerables and Ripley, 2002). RT is a nonlinear predictive model, implemented by recursive partitioning of the data and multiple model fitting for each partition (Strobl et al., 2009). The prediction of RT is sometimes unstable and sensitive to a small change in the tree structure. RF was improved from RT in order to increase the robustness of the model by assembling numerous trees and the prediction of RF was made by averaging a set of RTs rather than a single one (Breiman, 2001). DSM (RK with GLM, RT, and RF) was conducted by an automated fitting procedure of 'GSIF' package (Hengl, 2014) in R version 3.2.3.

Soil properties, including VWC, SOM, sand content, and clay content, were log-transformed and normalized before using in the DSM process following (Hengl et al., 2004):

$$X' = \ln\left(\frac{X}{1-X}\right); 0 < X < 1 \quad (4.4)$$

and back transformed after the prediction following (Diggle and Ribeiro, 2007):

$$X'' = \frac{1}{1+e^{X'}} - 0.5 \times V(X') \times e^{X'} \times (1 - e^{X'}) \times \frac{1}{(1+e^{X'})^3} \quad (4.5)$$

In addition, other soil properties, including GWC, EC, P, K, Na, Mg, Zn, and Mn were implemented with another log-transformation before using them in the DSM process following

$$X' = \ln(X); X > 0 \quad (4.6)$$

and back transformed after prediction following

$$X'' = e^{X'} \quad (4.7)$$

X is the original soil property within the range 0-1, X' is log-transformed soil property as input for DSM in equation (4.4) and (4.6) and output from DSM in equation (4.5) and (4.7), X'' is the back-transformed value from DSM prediction, and $V(X')$ is the sampling variance of the DSM prediction.

Equation (4.4) and (4.5) ensured that the predicted values were within the range 0-1 and equation (4.6) and (4.7) reduced the skewness of input data in order to satisfy a normal distribution and ensured that the predicted values were positive.

4.2.4 Validation

An internal uncertainty of DSM was expressed as standard error map calculated from the 103 samples during model fitting. In addition, an independent validation dataset with 45 soil cores was used to assess the prediction accuracy of 3D RK. Root mean squared error (RMSE) values of validation dataset were calculated for 10 soil depths of each soil property.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4.8)$$

where n was the number of samples, y_i was the measured value, and $\hat{y}_i$ was the predicted value.

4.3 Results and discussions

4.3.1 Descriptive statistics

Table 4.1 Descriptive statistics of predicted soil properties (calibration dataset in the upper table and validation dataset in the lower table) from vis-NIR spectra

Soil Properties	Units	Mean ± sd[1]	CV[2]	median	min	max	skewness	kurtosis
VWC	%	43.52±9.54	0.22	42.58	21.80	76.61	0.45	0.04
GWC	%	49.58±35.12	0.71	39.02	15.13	277.01	2.67	9.57
BD	g cm^{-3}	1.06±0.32	0.31	1.11	0.18	1.90	-0.32	-0.46
SOM	%	9.16±13.78	1.50	2.55	0.18	85.08	2.54	7.14
pH		7.15±0.52	0.07	7.13	5.70	8.48	0.09	-0.64
EC	μS cm^{-1}	326.33±257.26	0.79	256.24	71.79	2277.61	2.88	11.73
Sand	%	29.61±19.47	0.66	29.61	0.47	87.45	0.54	-0.51
Clay	%	43.55±20.21	0.46	38.42	7.28	97.85	0.63	-0.48
P	mg kg^{-1}	24.43±33.83	1.38	10.84	0.11	413.81	3.62	27.05
K	mg kg^{-1}	106.07±78.90	0.74	80.47	17.05	562.97	2.47	7.56
Na	mg kg^{-1}	68.83±50.86	0.74	57.16	23.77	859.78	7.25	86.28
Mg	mg kg^{-1}	643.01±343.38	0.53	539.52	139.05	2413.34	1.56	2.75
Zn	mg kg^{-1}	4.67±3.28	0.70	3.69	0.56	16.95	1.01	0.47
Mn	mg kg^{-1}	19.84±22.83	1.15	11.87	2.02	263.02	3.89	24.40
VWC	%	46.85±10.68	0.23	45.51	23.34	76.04	0.24	-0.68
GWC	%	65.68±53.39	0.81	45.91	15.42	357.68	2.18	5.80
BD	g cm^{-3}	0.97±0.37	0.38	1.01	0.17	1.82	-0.18	-0.90
SOM	%	11.49±16.39	1.43	2.60	0.31	81.88	1.93	3.50
pH		7.12±0.50	0.07	7.08	6.06	8.40	0.24	-0.79
EC	μS cm^{-1}	447.28±408.70	0.91	302.99	80.30	2875.66	2.58	8.73
Sand	%	25.54±18.92	0.74	20.26	0.52	80.73	0.78	-0.36
Clay	%	49.27±21.42	0.43	46.72	9.52	96.95	0.28	-0.93
P	mg kg^{-1}	23.14±27.09	1.17	11.20	0.10	137.30	1.52	1.83
K	mg kg^{-1}	113.19±88.38	0.78	81.27	20.45	611.90	2.41	7.45
Na	mg kg^{-1}	86.62±76.14	0.88	63.29	26.14	655.68	4.05	21.60
Mg	mg kg^{-1}	764.73±404.86	0.53	641.78	231.30	2384.34	1.10	0.89
Zn	mg kg^{-1}	5.75±4.06	0.71	4.54	0.66	17.15	0.72	-0.56
Mn	mg kg^{-1}	21.51±26.73	1.24	12.18	2.69	210.38	3.41	14.31

[1]sd: standard deviation; [2]CV: Coefficient of variation.

The basic statistics of calibration dataset and validation dataset were shown in Table 4.1. In calibration dataset, soil properties in this field are extremely variable with large coefficient of

variation. SOM, P, and Mn showed larger standard deviations than mean values. GWC, SOM, EC, P, K, Na, and Mn are highly positively skewed with skewness of 2.67, 2.54, 2.88, 3.62, 2.47, 7.25, and 3.89. BD is slightly negatively skewed with skewness of -0.32. Most soil properties in validation dataset showed similar ranges as in calibration dataset, such as VWC, BD, SOM, pH, sand content, clay content, and Zn. Therefore, the validation dataset are fully representative and could provide enough power for model evaluation. However, P, Na, and Mn had smaller ranges in the validation dataset compared to the ranges in calibration dataset. GWC, EC, and K showed larger ranges in validation dataset compared to the ranges in calibration dataset, thus requiring high extrapolation ability from mapping process.

4.3.2 Comparison of GLM, RT, and RF

RK with PCA effectively interpreted the soil-landscape relationship, displayed the spatial pattern of variability, and captured the detailed variation in soil properties. It has been reported to outperform OK by involving effective regression techniques (Levi and Rasmussen, 2014; Zhu and Lin, 2010) while some reported no significant improvement of RK over other pure regression or kriging methods (Li, 2010; Vaysse and Lagacherie, 2015). The performance of RK was also dependent on a strong relationship between soil properties and auxiliary variables (Sun et al., 2012a). Additionally, RK increased the normality of residuals (Hengl et al., 2004), thus satisfying the input criteria for interpolation method compared to initially skewed distribution of soil properties.

In this study, the combination of RK with a linear model (GLM) and two non-linear models (RT and RF) were compared for the prediction accuracy and efficiency. All the models produced similar patterns of the spatial variability of soil properties. However, differences were observed between these models when we considered the fitting process, variograms, fitting accuracy, and final map products. GLM lies on a linear relationship between soil properties and environmental variables which increased its limitation on explaining the complex soil-landscape relationship compared to non-linear models. Moreover, RF, aggregating a set of RT models, increased the stability and robustness over a single RT model. During the fitting process, even though the regression model residuals were fitted with OK, they were expected to have low spatial dependency reflected by shorter range and bounded sill of variograms (Hengl et al., 2004), and this was better shown in the RF model (Fig. 4.3). However, due to the weak spatial structure of

residuals in RF model, the variogram was fitted by increasing the range which resulted an inconsistent fitting graph and point pattern.

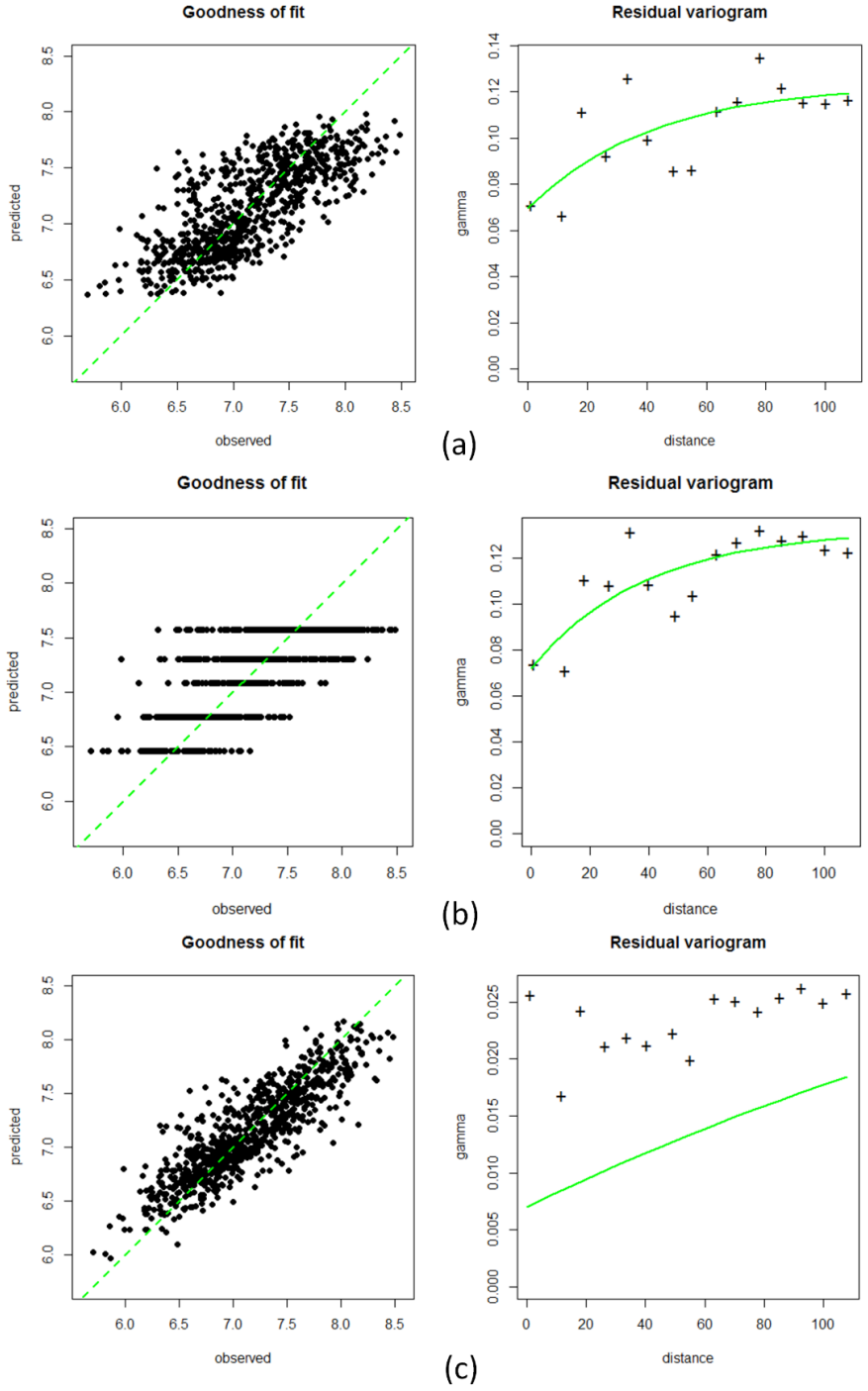

Fig. 4.3. Fitting results (goodness of fit and residual variogram) of RK with three models; a) GLM, b) RT, and c) RF in 3D-DSM of pH. The dash lines in the left column indicated 1:1 line. The dashed lines in the right column indicated variograms.

The goodness of fit (Fig. 4.3) showed relatively scattered points of GLM from 1:1 line compared to RF hence indicating worse fitting results. RT is a recursive partitioning process and resulted in segmented features in fitting results (Fig. 4.3). The RMSE values in the independent validation dataset and their changes with depth were shown in Fig. 4.4. In general, there are no significant changes of values and distributions of three models for BD, EC, K, and Mg in the whole profile. RF significantly outperformed GLM for sand content, clay content, and P with the average RMSE values decreased by 9%, 15%, and 30%, respectively. In addition, RF outperformed GLM for VWC only in deep soil with the average RMSE value decreased by about 10%. Furthermore, the average RMSE value decreased by about 10% by RF compared to GLM and RT in the whole profile for pH and Mn. However, RF resulted in worse results for GWC, SOM, and Na as the average RMSE values of the whole profile increased by more than 10% in comparison with GLM and RT.

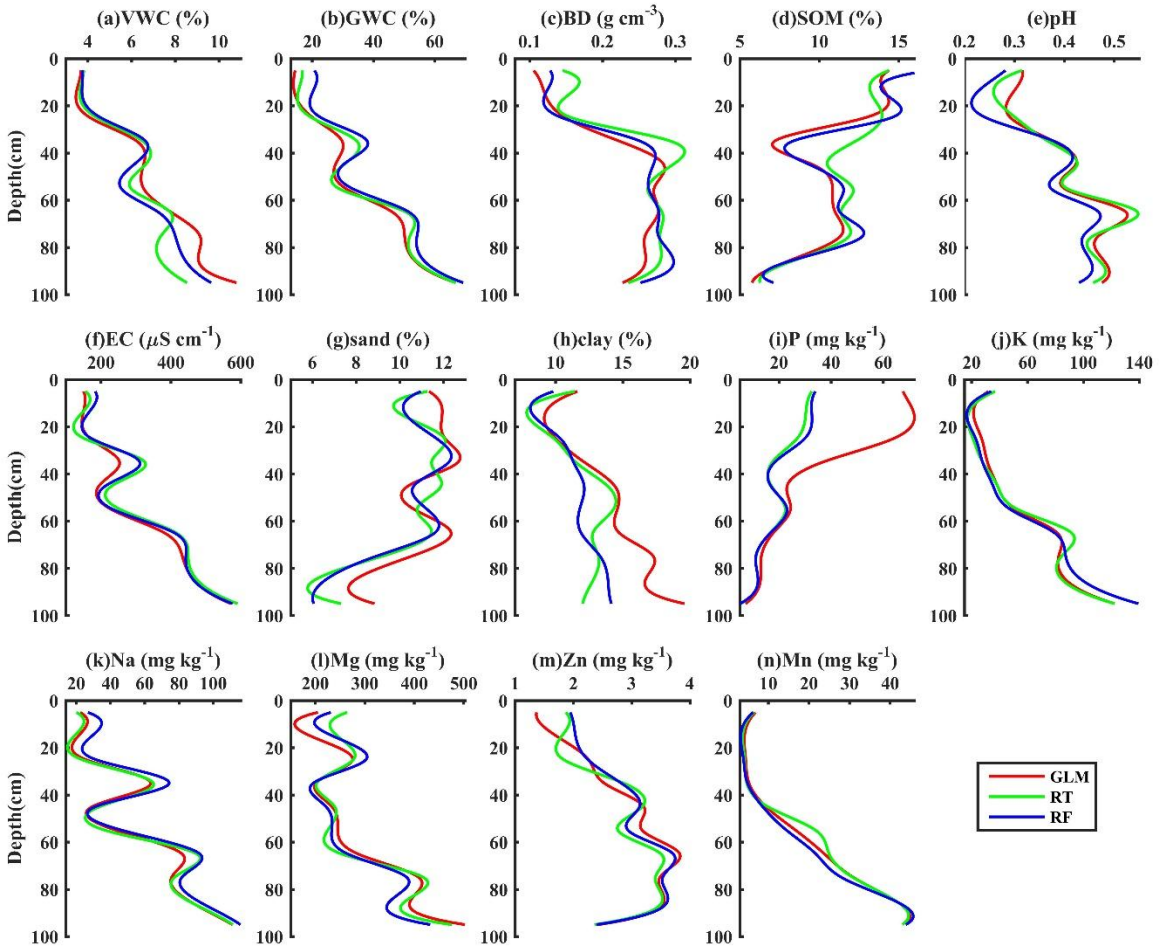

Fig. 4.4. Prediction results (RMSE) with depth of the independent validation dataset by GLM, RT, and RF models in DSM of soil properties.

Even though the spatial pattern were similar in the final maps by GLM, RT, and RF, the maps produced by RT displayed unsmooth changes and segmented features due to the partitioning process of model fitting (Fig. 4.5, 4.7, 4.9). RF was visually considered to capture the fine details of variation by breaking the big patches of the same color in the maps produced by GLM. The predicted values generally maintained the similar ranges of measured soil properties by GLM model, whereas RT and RF models smoothed the variability and narrowed down the ranges for many soil properties, including GWC, SOM, EC, P, K, Mg, and Zn. For example, SOM in calibration dataset ranged from 0.18% to 85% and predicted values of RK by GLM, RT, and RF ranged from 0.27% to 91%, 0.37% to 74%, and 0.35% to 62%, respectively (Fig. 4.7). This was consistent with the results of Schmidt et al. (2014) who reported a wider range of sand content obtained by linear model (MLR) compared to the RF model. They also argued that the accuracy shown by a wider range and details in linear model was spurious as it actually did not increase the accuracy of final maps (Schmidt et al., 2014), and even worse (larger RMSE and sampling variances compared to RF) in our study. GLM could not maintain the similar range when the range was long and the distribution was skewed. For example, the range of Na in calibration dataset was from 24 to 859 mg kg^{-1}, whereas the ranges of predicted values were from 26 to 311 mg kg^{-1}, 29 to 294 mg kg^{-1}, and 25 to 306 mg kg^{-1}, by GLM, RT, and RF models, respectively. Similar result was obtained for Mn. In addition, the range of P was 0.11 to 414 mg kg^{-1} in the calibration dataset, while the predicted range of P was 0.33 to 1540 mg kg^{-1} by GLM. This showed that GLM is prone to be influenced by small interference from covariates. Such interference was also shown in the uncertainty maps. As an index of the prediction uncertainty, the standard error maps of corresponding soil properties provided information on the confidence that one could have on the prediction results. By comparison, RF models with smallest standard errors for all the soil properties and all the layers showed more confident prediction and narrower confidence intervals. Generally, GLM and RT resulted in large and similar standard errors for all the soil properties. While GLM resulted in slightly larger standard errors in BD, sand, clay, K, Na, Mg, and Mn in deep soil, and RT resulted in slightly larger standard errors in Zn in deep soil. In addition, the several abrupt red points in the bottom left of GLM standard error maps (Fig. 4.6, 4.8, 4.10) showed extremely high standard errors which might have caused by noise features in environmental covariates, while RF model had the advantages of omitting the noise features and avoiding over-fitting problem (Grimm et al., 2008).

Therefore, by taking into account all the factors (fitting processes, variograms, validation results, and map uncertainty) associated with three regression models in RK, RF model outperformed GLM and RT in this study by interpreting a complex non-linear relationship between soil-landscape relationships, implementing a robust fitting process, capturing a fine variation in soil properties, and resulting in higher prediction accuracy with lower RMSE and sampling variance. The digital soil maps from 3D RK with RF model will be further used to understand the spatial variability of multiple soil properties.

4.3.3 Map products (RF)

4.3.3.1 Spatial distribution

As important fundamental constituents of soil, SOM, sand, and clay play an essential role in determining spatial patterns of many other soil properties. Predicted values of SOM ranged from 0 to 62%, narrower in comparison to values in calibration dataset that ranged from 0 to 85%. For the top soil (0-10 cm depth), the highest SOM content was observed at the bottom left of the field, which was mainly dominated by organic soil with SOM content of 50% to 60%, and the lowest SOM was observed in the middle left of the field, the area covered by fine sandy soil with comparatively very low SOM content (3% to 4%) (Fig. 4.7). The upper part of the field was dominated by shallow organic deposits and the middle right part of the field was dominated by deep organic deposits and exhibited higher SOM content of about 30% to 40%. SOM content dramatically decreased with depth, and the decreasing pattern was quantified by an exponential decay function (Minasny et al., 2006). The variability in the sub and deep soil was highly skewed with about 30% SOM content in the organic soil profiles located in the bottom left of the field, while with SOM content less than 5% for the rest of the field. The predicted values of sand content ranged from 1% to 78%, while the original range was between 0 and 87%. The distribution pattern of sand content was similar but opposite to the distribution pattern of SOM in the top soil (Figures not shown in this thesis). The higher the SOM, the lower the sand content, and vice versa. Moreover, the sand content significantly increased from top soil to sub soil in the sandy soil region (middle left of the field) and slightly decreased thereafter in the deep soil (sand content higher than 65%). In general, the coarse sand replaced the fine sand in the sub soil. In the other region of the field, sand content significantly and continuously decreased with depth. This may be attributed to the formation of landscape in the area and the presence of Champlain Sea in this region. The range for predicted clay content was from 9% to 95% and was close to the range of original clay content

(between 7% and 98%). The distribution of clay content was similar to SOM and sand content in the top five layers. The region with high SOM content had high clay content and the region with high sand content had low clay content. Unlike SOM, the clay content continuously increased with depth throughout the field but the horizontal pattern did not change too much from top to the deep soil layers. However, a drastic increase in the clay content was observed in the upper part of the field.

Almost all the other soil properties were more or less affected by the SOM, sand, and clay content, which was also reflected in their spatial distribution. SOM predominantly affected other soil properties in the top soil and organic soil region with high SOM content. SOM is significant for its high capacity of holding water and cations, increasing the stability of soil aggregates, building a good soil structure, and having large specific surface area and porosity (Bot and Benites, 2005). Sand content played an important role in influencing soil properties in the sandy region of the study area with fine sand in the top soil and coarse sand in the sub soil. Sandy soil with large pores and low specific surface area has low capacity to hold water and cations (Bruand et al., 2005). On the other hand, clay has strong ability to hold water and base cations owing to the fine pore size with large amount of total pore, and large specific surface area and mainly affected soil water content, EC, CEC, and soil cations in the deep soil with high clay content in this study (Pask and Turner, 1955).

The range of predicted values of VWC was between 24% and 70%, slightly narrower than the original range between 21% and 77%. The horizontal distribution of VWC were quite similar to that of SOM and clay content (Fig. 5.5). SOM mainly accounted for the high VWC in the top soil while clay was responsible for the high VWC in the deep soil. In general, the VWC content increased with depth and clay content except at the middle left region of the field which was dominated by sandy soil (Fig. 5.5). The relatively low VWC decreased at the subsoil (30-60cm) due to the change of soil texture from fine sand to coarse sand and increased thereafter from increasing clay content. The range (17% to 209%) of predicted values of GWC was significantly smaller than original range between 15% and 277%. GWC showed a similar spatial variability pattern as of VWC in the top soil and sub soils (Figures not shown). However, no increase in GWC was observed in deep soil with increasing clay content. GWC is a mass-based water content index and takes the soil weight into account (Lambe and Whitman, 1969). In spite of strong ability to store water, the increasing soil weight (used as denominator in the calculation) reduced the GWC.

Therefore, the distribution of GWC was more like the distribution of SOM. The predicted values of BD ranged from 0.31 to 1.72 g cm^{-3}, slightly different from the original values within the range of 0.18 and 1.90 g cm^{-3}. The spatial variability of BD in the top three layers was similar to that of SOM with opposite magnitude as high SOM content corresponded to low BD and low SOM content (higher sand content) corresponded to high BD (Figures not shown). The loose organic material corresponded to low BD in the bottom left organic soil region. With increasing depth, the BD increased as the SOM decreased and the mineral matter (clay) increased. However, in the sandy soil region, the BD increased drastically from the top soil to the sub soil with the increase of coarse sand. Furthermore, the decrease in BD was attributed to the increase in clay content, while still higher than other regions dominant by clay soil.

The range of predicted pH was between 5.85 and 8.18, while the range of original pH was between 5.70 and 8.48. Though a general spatial structure was observed, it was not quite similar and strong as of other soil properties (Fig. 4.9). Generally, soil pH is an outcome of an indirect effect of many soil properties (with major influence from SOM and soil cations) rather than direct effect from one or two soil properties. In addition, the small-scale variation was clear with abrupt changes and was indicated by a relatively large nugget of the variogram. The small region in the middle left of the field with sandy soil had lowest soil pH, and the organic soil region in the bottom left of the field had relatively lower soil pH as well. Low cation exchange capacity (CEC) of sandy soil attributed to lower soil pH while the organic acids released during decomposition of SOM lowered the soil pH in organic soil. Soil pH gradually increased with depth due to the decrease of SOM and the increase of clay content with higher CEC. A similar pH value was observed for the bottom 4 layers (60 cm -100 cm) may be due to the absence of roots and greater effect of shallow ground water. The predicted values of EC ranged from 82 to 1472 μS cm^{-1}, significantly smaller than original values with range between 72 and 2278 μS cm^{-1}. Soil EC measures the amount of salts in the soil and is affected by many soil properties including soil texture, SOM, and CEC (Grisso et al., 2009). The distribution of EC in the top soil layers were similar to the distribution of SOM with high EC in the bottom left and upper part of the field and low EC in the sandy soil region (Figures not shown). With increasing depth, EC decreased in the organic soil region with decreasing SOM content, and increased in the upper right of the field with increasing clay content. Clay soils with high CEC generally result in higher EC (Grisso et al., 2009).

While the spatial distribution of SOM and pH mainly contributed to the distribution of P, clay contributed to the distribution of cations including K, Na, Mg, Zn, and Mn (Figures not shown in this thesis). The range of the original values of P was between 0.11 and 414 mg kg^{-1} and the range of the predicted values was between 0.46 and 257 mg kg^{-1}. Major variation was observed mainly in the top three layers where the values ranged between 12 and 257 mg kg^{-1} while the values ranged between 0.46 to 62 mg kg^{-1} in other layers. Similar to the distribution of pH, high values of P in the upper layers were mainly concentrated in the sandy soil region and showed a strong negative correlation (0.67) with pH. K and Mn showed similar horizontal and vertical distribution. High values of these cations were located in the organic soil region and low values were located in the sandy soil region as observed from their horizontal distribution. The values of these cations increased with depth as the clay content increased. The most significant and drastic increase of these cations were observed mainly in the upper region. The predicted values of K ranged between 25 and 477 mg kg^{-1} while the original values ranged between 17 and 563 mg kg^{-1}. The predicted values of Mn ranged from 4 to 119 mg kg^{-1} and original values of Mn ranged from 2 to 263 mg kg^{-1}. The ranges of predicted values were narrower for these cations than that of the measured values and this difference was attributed to the influence of log-transformation that could not accurately predict long tail in the relatively high values. The predicted values of Na ranged from 25 to 306 mg kg^{-1} and the original values ranged from 24 to 860 mg kg^{-1}. The locations with high Na located in the bottom left part of the field with high SOM. A slightly increasing amount of Na was observed with the increase of clay amount from top soil to sub soil. In addition, in the sandy soil region, the concentration of Na slightly increased with depth while remained unchanged in other regions. The predicted values of Mg ranged from 162 to 1701 mg kg^{-1} and the original values ranged from 139 to 2413 mg kg^{-1}. Mg was influenced by both SOM and clay content and had a spatial distribution similar to that of VWC. The predicted values of Zn ranged from 0.71 to 13.74 mg kg^{-1}, similarly compared to the measured values (0.56 to 16.95 mg kg^{-1}).The concentration of Zn was also influenced by both SOM and clay content. A high amount of Zn concentration corresponded to high amount of SOM in the surface layers and gradually decreased with depth.

4.3.3.2 Uncertainty maps

The standard errors of all the soil properties showed almost exactly same trend. The standard errors slightly decreased from the top soil to the sub soil but without any significant difference (Fig. 4.6, 4.8, 4.10). A smaller number of samples in the surface layers might have increased the standard

error. This may be attributed to contact issues between vis-NIR spectra probe and organic soil (loosened) in the top layers leading to the loss of spectra data (about 3). From the spatial distribution of standard errors mainly at the bottom four layers, a large standard error was observed close to the right side of the field where collected soil cores were mainly shallow (only to a depth of about 40 to 50 cm). The standard error is an index of prediction uncertainty, which is primarily determined by the available information provided to the model. The more the information available (larger number of sampling points), the smaller the standard error and the greater the confidence of the prediction. Therefore, this significant decrease in sampling points and thus lower amount of information available for model prediction increased the standard error in right side of the field in deep soil.

4.3.3.3 Validation results

In addition, RMSE values of an independent dataset was used as a direct index of prediction accuracy (Fig. 4.4), which is mainly influenced by the complexity of the environment, the spatial variability of soil properties, and the predictive ability of models. The RMSE values of K and Mn showed monotonically increasing trend with depth, and this might be due to the increasing magnitude of actual values of K and Mn from top soil to deep soil. A fluctuating increase of RMSE values was observed for majority of soil properties, including VWC, GWC, BD, pH, EC, clay content, Na, Mg, and Zn, whereas a fluctuating decrease was observed for SOM, sand content, and P. The top three soil layers were mainly plow layers which were highly affected by the mixed effects of natural processes and imposed agricultural activities. The mixing effects of tillage generally produced relatively uniform conditions within the plow layers and were also reported by many studies (Kempen et al., 2011; Liu et al., 2016). Moreover, most of the environmental variables used to build the calibration relationship with soil properties were collected from the surface. These two factors contributed to higher accuracy in the top soil. Generally, due to less available environmental variables in deep soil, the RMSE values were largest for majority of soil properties. Sub soil is a transition from the bottom of plow layers to the bottom of root zone, and environment is complex in such region, thus resulting in fluctuating changes of RMSE values. However, for SOM, P, and sand content, the largest variation mainly happened in top soil, and simplest environment and the smallest spatial variation existed in deep soil. The variation of SOM in the deep soil was almost half of the variation in the top soil. This variation was so simple and

highly skewed as large number of samples were observed with low SOM content in the deep soil. Therefore, the RMSE values of SOM, P, and sand content gradually decreased with depth.

Comparing current literature of 3D-DSM products, different results were obtained as for the different performance of prediction accuracy with depth. Higher accuracy of top soil prediction than sub soil was also reported by Piikki et al. (2015) in 3D mapping of sand and clay. However, Lacoste et al. (2014) reported a high accuracy in sub soil layers (15-60cm) in 3D mapping of SOC. Kempen et al. (2011) demonstrated the highest prediction accuracy in the top soil and lowest prediction accuracy in the sub soil for 3D mapping the SOC content using depth functions. A more comparable result was reported by Vaysse and Lagacherie (2015) by using RK-RF model. The RMSE values of clay content and SOC showed similar trends: a slight increase with depth and a decrease thereafter. However, sand content and pH showed generally increasing RMSE values with depth (Vaysse and Lagacherie, 2015). Taghizadeh-Mehrjardi et al. (2016) observed an increasing RMSE values with depth and a higher accuracy in the soil surface for 3D mapping of SOC. Above all, higher accuracy was mainly observed in top soil as the environmental covariates were collected in soil surface and the calibration relationship was stronger in top soil. Some soil properties (especially SOM) was lowest in deep soil. Therefore, the smaller variation and simple environment in deep soil contributed to higher accuracy in some cases. While in most cases, due to the complex and unknown environment in deep soil, the accuracy is always lower. Sub soil is the trickiest horizons with either good prediction or poor prediction, and further exploration is required for understanding the complex subsoil environment and improving the prediction accuracy in the sub soil. In addition, more studies should pay attention to deep soils due to the enormous importance of whole soil profile variability information and its special essence for deep soil interpretation and management e.g. soil drainage, carbon storage, and leaching dynamics.

4.4 Conclusions

Three-dimensional digital soil maps were prepared for a large number of soil properties by regression-kriging. Three regression techniques, including generalized linear model, regression tree, and random forest were tested and compared to identify the most effective prediction with regression-kriging. In addition, proximal soil sensing techniques (vis-NIR spectra and gamma-ray radiation) was used to densely collect soil information for input of DSM and used as environmental covariates. The results were presented through a series of final map products as well as the associated standard error maps for multiple soil properties.

In brief, soil vis-NIR spectra can exhaustively and accurately obtain soil property information and provide sufficient input for DSM. As for three models in the DSM process, the RF model showed the advantages of interpreting non-linear soil-landscape relationship, fitting weak spatial dependency of regression residuals, and resulting in higher validation accuracy and smaller prediction uncertainty. Therefore, it was regarded as a superior model for RK over GLM and RT. By interpreting the spatial variability and prediction accuracy of final maps, SOM, sand, and clay showed clear horizontal and vertical distribution and contributed greatly to the spatial distributions of other soil properties. A high SOM was observed at the bottom left part of the field with soil series of deep organic deposit and upper part of the field with soil series of shallow organic deposit. SOM also significantly decreased with depth. Clay content exhibited a similar horizontal distribution of SOM but greatly increased with depth. The ability of SOM and clay to hold water and cations played an essential role in the distributions of other associated soil properties (water content, pH, EC, P, and soil cations).

The mapping uncertainty expressed by standard errors was mainly determined by the sample size so that it displayed similar trends: standard error decreased slightly from the top soil to the sub soil with the largest values at the deep soil especially in the middle right part of the field. The validation accuracy quantified by RMSE values of an independent validation dataset showed that for majority of soil properties, largest accuracy obtained in soil surface due to the uniform environment in the plow layer and sufficient environmental covariates collected in the soil surface. The accuracy gradually decreased with depth due to large values of many soil properties and the complex environment in deep soil. However, SOM, P and sand content showed opposite distribution due to the decreasing trend of the actual values with depth.

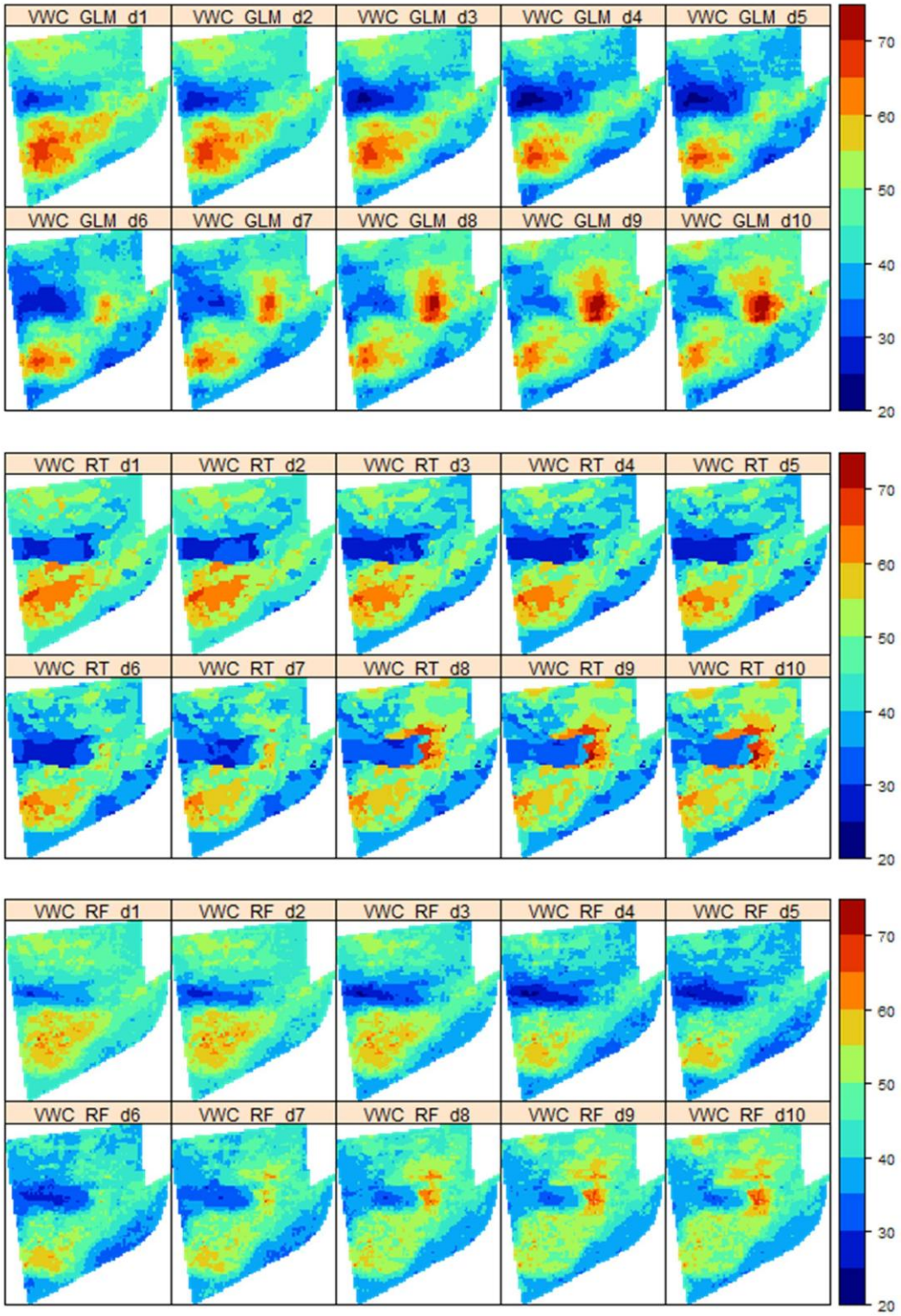

Fig. 4.5. Maps of volumetric water content (VWC) at different depths (d) produced using generalized linear model (GLM), regression tree (RT), and random forest (RF).

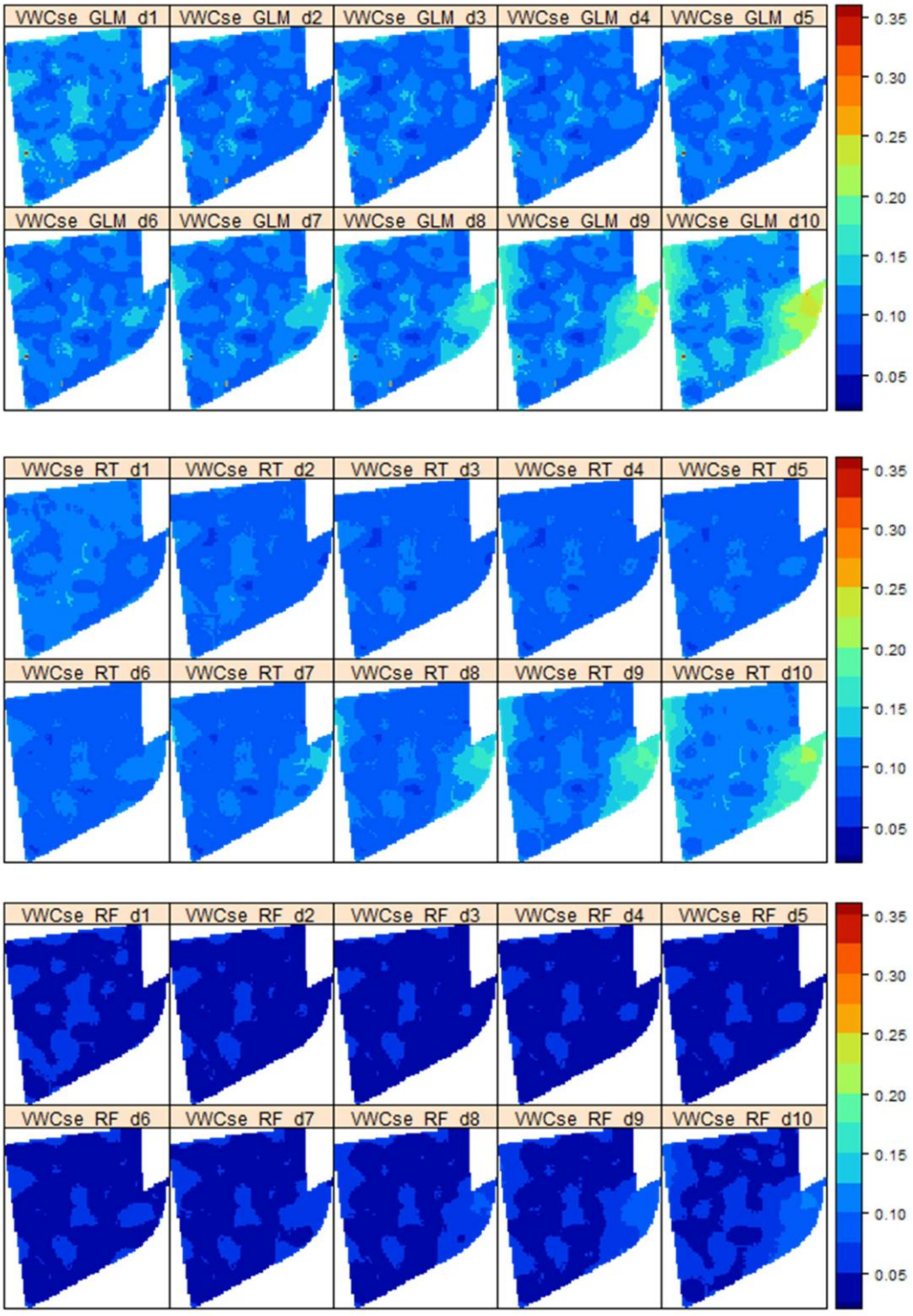

Fig. 4.6. Standard error maps of volumetric water content (VWCse) at different depths (d) produced using generalized linear model (GLM), regression tree (RT), and random forest (RF).

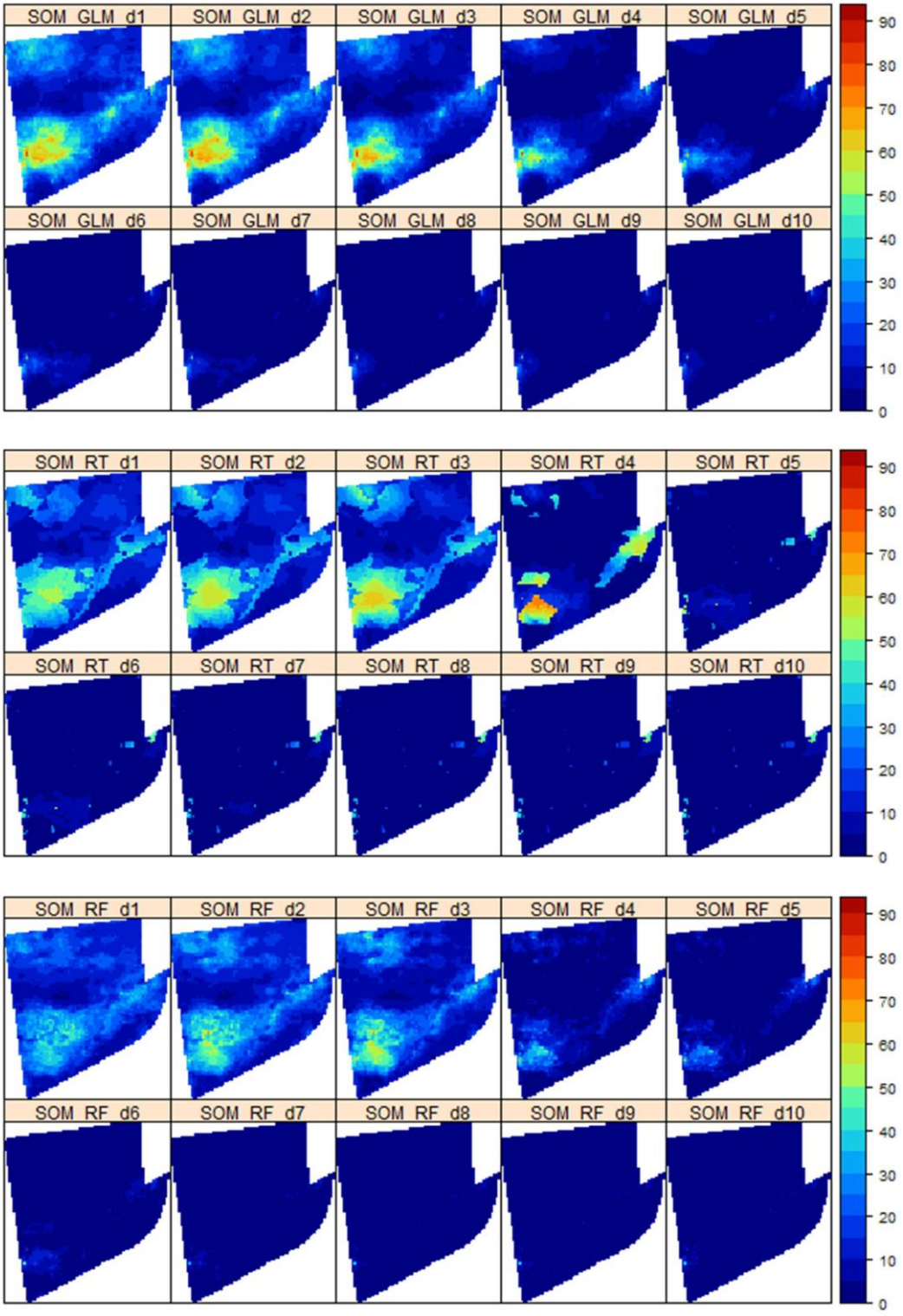

Fig. 4.7. Maps of soil organic matter (SOM) at different depths (d) produced using generalized linear model (GLM), regression tree (RT), and random forest (RF).

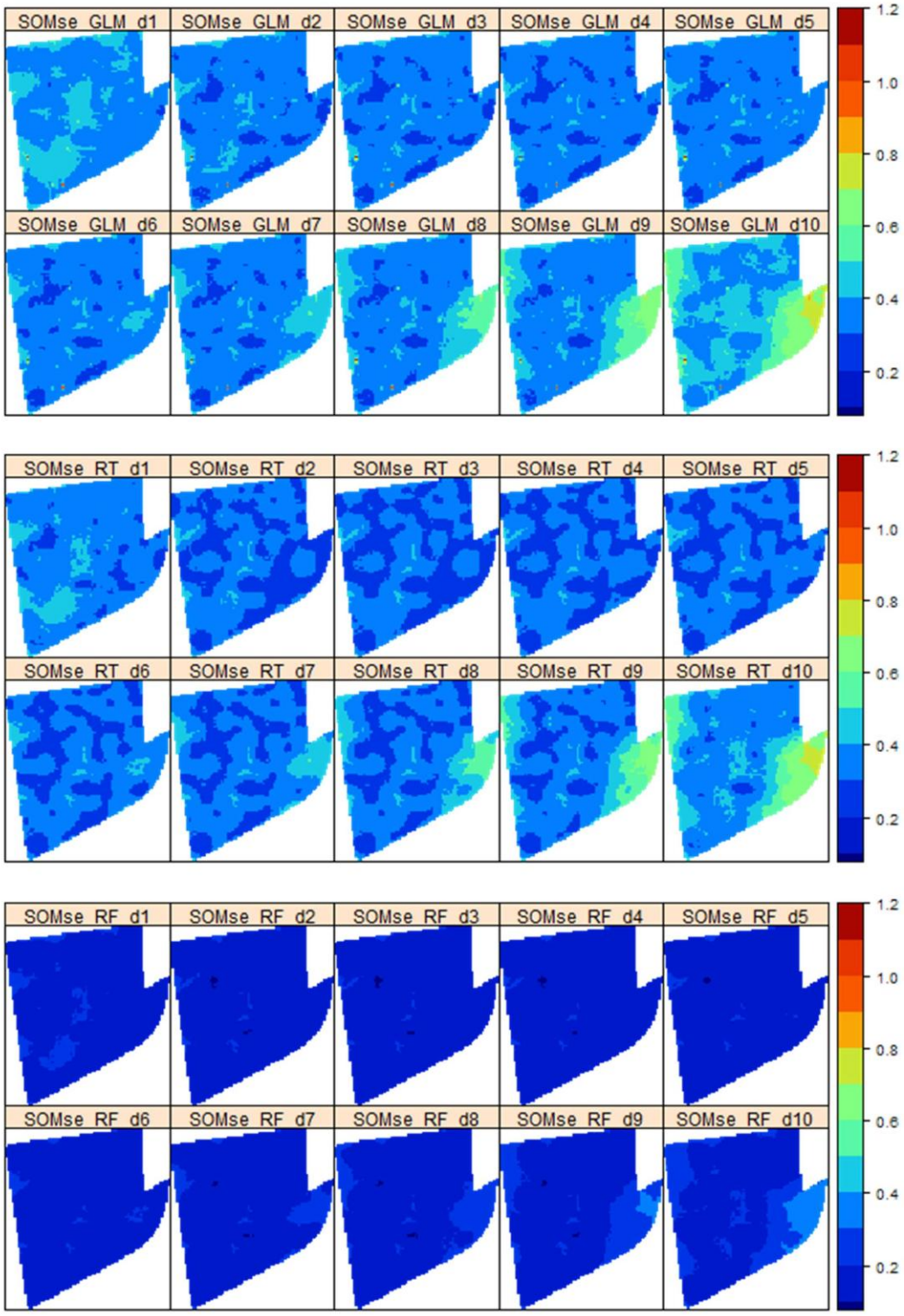

Fig. 4.8. Standard error maps of soil organic matter (SOMse) at different depths (d) produced using generalized linear model (GLM), regression tree (RT), and random forest (RF).

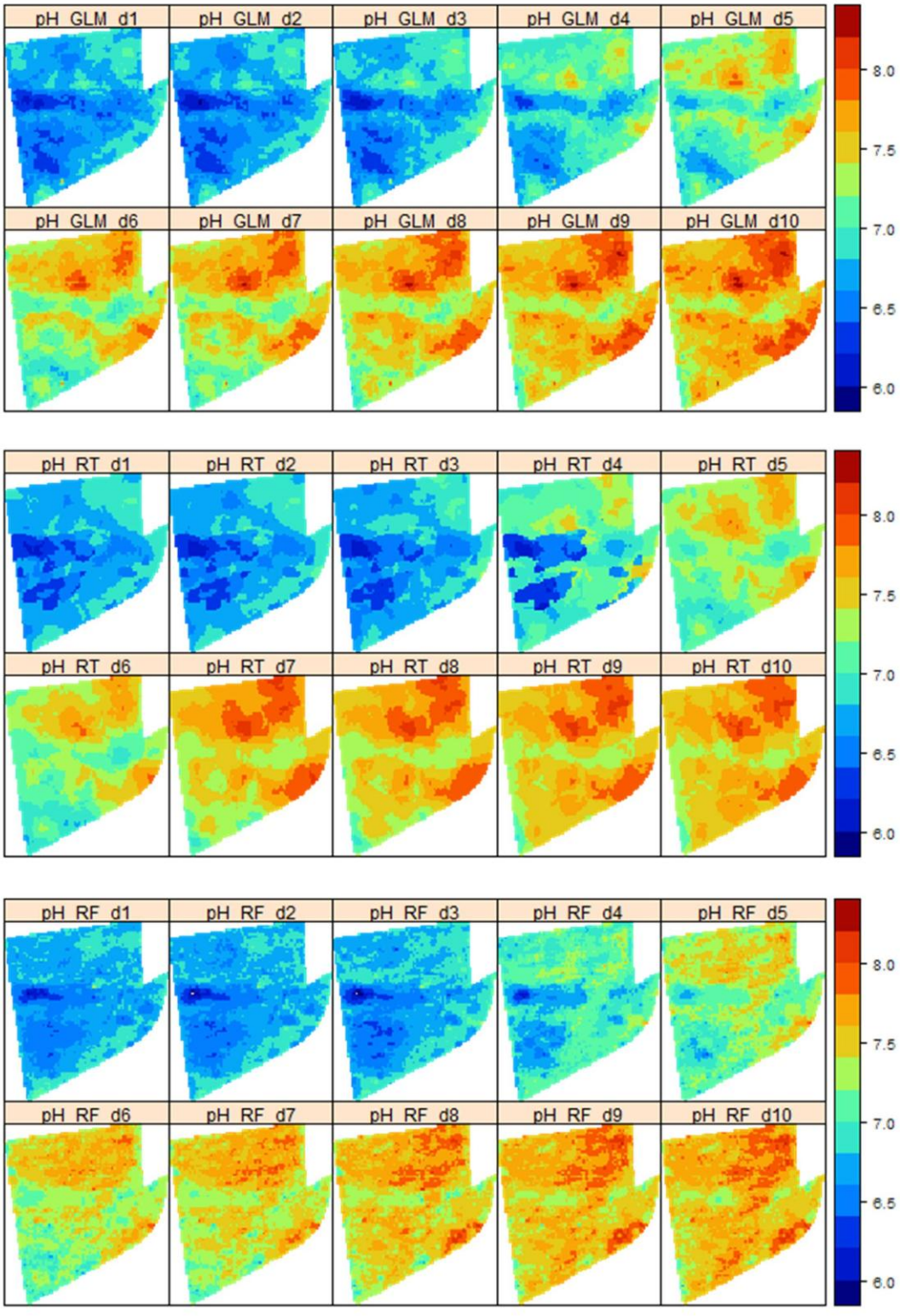

Fig. 4.9. Maps of soil pH (pH) at different depths (d) produced using generalized linear model (GLM), regression tree (RT), and random forest (RF).

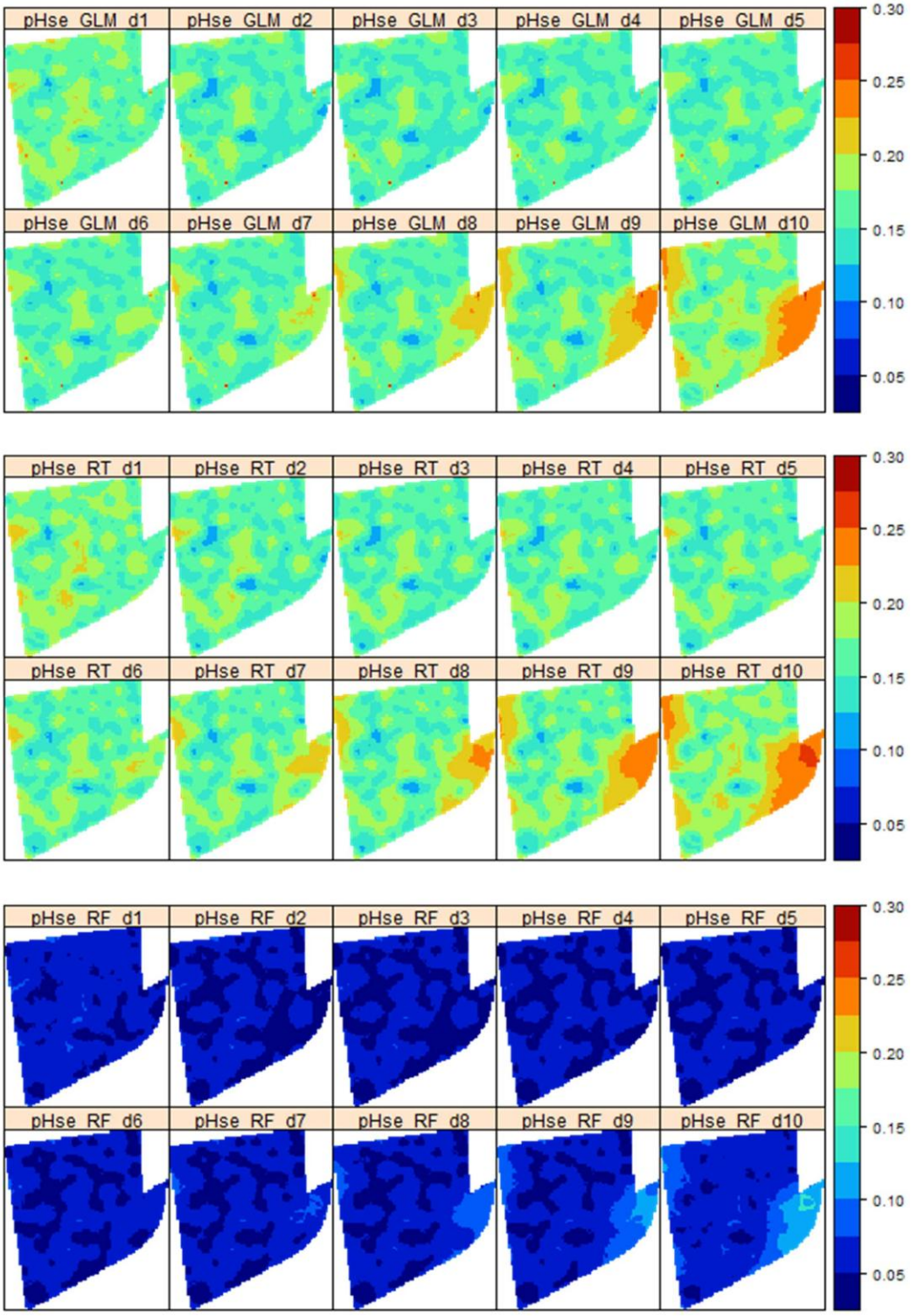

Fig. 4.10. Standard error maps of soil pH (pHse) at different depths (d) produced using generalized linear model (GLM), regression tree (RT), and random forest (RF).

PREFACE TO CHAPTER 5

In Chapter 4, three different regression techniques (generalized linear model, regression tree, and random forest) recognizing the linear or non-linear relationships between soil properties and environmental covariates were initially compared to select the method that was the most effective for use with regression kriging. As random forest method was finally selected as the best model, 3D digital soil maps were developed for multiple soil properties from total 148 soil profiles by 3D regression kriging with the random forest method. These maps were regarded as original maps. As the first step of digital soil mapping, sampling design plays an essential role in selecting and providing reliable soil data into DSM. An optimized sampling design uses a small sample size with specific selection criteria for DSM thereby greatly reduces the effort without compromising the mapping accuracy. Therefore, the main purpose of Chapter 5 is to compare and identify the optimized sampling design in order to make the DSM more efficient. Six SDs (grid sampling, grid random sampling, simple random sampling, stratified random sampling, transect sampling, and conditioned Latin hypercube sampling) were used to select a small sample size (45 soil profiles) for DSM. These sampling designs were chosen after reviewing the recent literatures in Chapter 2 as these were commonly used sampling designs in DSM optimizing the sample patterns either in geographical space or feature space. These maps were regarded as sub maps. The sub maps were produced by the same procedure as original maps produced in Chapter 4 and were further compared with original maps to identify the one that had higher accuracy and better reproduced the original maps. Two optimization criteria (geographical space and feature space coverage) were used to assist the comparison of sampling designs. Additionally, individual independent dataset with 103 soil profiles of corresponding SDs was used to validate the accuracy of sub maps.

The Chapter 5 has been written as a research paper format and will be submitted to *Geoderma* (impact factor 2.85). The detailed information for authors is shown below:

Order of authors: Yakun Zhang[a], Asim Biswas[a,*], Wenjun Ji[b], & Viacheslav I. Adamchuk[b].

The contribution of authors are the same as Chapter 3 and 4. The author of this thesis took charge of all the lab work, data analysis, and manuscript writing. Prof. Biswas as thesis supervisor was completely involved in all stages of this study. Wenjun helped with the vis-NIR spectra data preparation and advised on the spectral model development. Prof. Adamchuk as thesis committee provided much support on the technical facilities, environmental variables collection and compilation, data analysis, and much suggestions on sampling design analysis.

CHAPTER 5

Comparison of sampling designs for calibrating three-dimensional digital soil maps

Abstract:

Incorporating steps of sample collection, model calibration, and validation, digital soil mapping (DSM) aims to produce elaborate maps of soil properties or soil classes for improved agricultural management and soil quality assessment. As an integral component of DSM, optimized soil sampling is essential for a reliable calibration model. With emerging 3D-DSM, soil profiling is needed and this may incur substantial costs. Therefore, the purpose of this study was to compare several common sampling designs, including: 1) grid sampling (GS), 2) grid random sampling (GRS), 3) simple random sampling (SRS), 4) stratified random sampling (StRS), 5) transect sampling (TS), and 6) conditioned Latin hypercube sampling (cLHS) for calibrating 3D-DSM models. The sample size of each sampling design is 45 samples. A field experiment was conducted at an agricultural field of Macdonald Campus Farm, Ste-Anne-de-Bellevue, QC, Canada. A total of 148 locations were identified by incorporating all the six sampling designs. Soil vis-NIR spectra data were collected at these 148 locations down to about 1 m depth using the Veris® P4000 soil profiler. In addition, a subset of 32 locations was used to collect soil cores to about 1 m and then, subdivided at 10 cm depth intervals. Thus, a total of 251 samples were analyzed in the laboratory for a range of soil physical and chemical properties (VWC, GWC, BD, SOM, soil pH, EC, sand content, clay content, available P, K, Na, Mg, Zn, and Mn). PLSR models were used to calibrate the relationship of vis-NIR spectra and soil properties at 32 locations, and then the relationship was used to predict the soil properties at the remaining 116 locations. Soil properties at 148 locations were used for 3D-DSM (original map) by regression kriging with random forest model. Additionally, soil properties were also mapped (sub maps) following the same procedure using samples collected at 45 locations following different sampling designs. The spatial distribution and uncertainty of each sub map were compared with the original map. Spatial and feature space coverage of sampling designs were also compared for six sampling designs. Furthermore, the rest 103 locations corresponding to every sampling design were used as validation datasets to evaluate the mapping accuracy. Results showed the strong influence of sampling design on the accuracy of digital soil maps. In general, stratified random sampling better represented the distribution of original maps and showed smaller RMSE values. While smaller RMSE values in the validation

dataset were observed in cLHS and SRS. Feature space coverage showed more essential effect on the accuracy of a specific sampling design over spatial coverage.

5.1 Introduction

Incorporating multiple steps from initial sample collection to model prediction and mapping to the validation of maps, digital soil mapping (DSM) is a complex project and requires elaborate designs for every integral step. Sampling design (SD), the first step of the whole mapping procedure, is essential for providing a reliable input for the calibration model. A sound SD makes a great difference for the subsequent laboratory analysis and statistical models. On the contrary, despite a good predictive model, it cannot compensate bad results obtained by a poor SD (Bui et al., 2006). A SD consists of two essential components: sample size and sample locations (Brus and Heuvelink, 2007). Sample size is determined by controlling the trade-off between budget and accurate information on landscape variability. It is intuitive that a large sample size can better reflect the variability of soil properties. However, money and labor costs associated with a large sample size are always beyond what one can afford (Kerry and Oliver, 2007; Kosmelj et al., 2001). Therefore, an optimized sampling size should be cautiously decided that makes the soil survey as cheap as possible by minimizing sample size, while simultaneously providing accurate information on spatial variation for prediction (Brungard and Boettinger, 2010). Sample locations are determined by different SDs. Sampling design is not the actual set of sample locations, but the procedure used to select it (Brus and de Gruijter, 1997). Sampling design is a systematic and complex science based on either rigorous derivations of statistics or auxiliary information about the variability of environment (Zhang and Zhang, 2011). With the purpose of providing reliable input for calibration model, SDs in DSM are usually optimized by either providing a good coverage in geographical space or feature space (Minasny and McBratney, 2007). Feature space, also called state space, attribute space, is a virtual space consisted of a set of environmental covariates (Hengl et al., 2003). A good coverage of feature space ensures a full representation of expected soil properties by environmental covariates, so that the prediction model will not be required to extrapolate beyond its bounds (Minasny and McBratney, 2007). Effective SDs that simultaneously optimize the sampling locations in geographical space and feature space has been explored (Hengl et al., 2003).

Three-dimensional digital soil mapping (3D-DSM) has become popular for its ability to interpret both the horizontal and the vertical variability of soil properties. Various SDs have been used in

guiding a sample collection for the 3D-DSM. For example, grid sampling (GS) with relatively uniform spatial coverage has been widely used in DSM (Veronesi et al., 2012). A modified GS with triangular grid was used by (Michot et al., 2013) for EC mapping at three soil layers, and by Malone et al. (2011) for 3D SOC mapping with profile depth functions. Stratified random sampling (StRS) that utilizes available soil or environment information of the study area to optimize the sample pattern was used by Vasques et al. (2010) for mapping SOC in multiple depths. Purposive sampling (PS) with the assistance of proximal sensing data was also applied by Huang et al. (2015) and Huang et al. (2014b) to identify the sensitive points used for 3D-DSM of soil salinity and particle size fractions, respectively. Furthermore, conditioned Latin hypercube sampling (cLHS) with the ability of obtaining a full coverage of multivariate distribution has been commonly used in 3D-DSM (Lacoste et al., 2014; Taghizadeh-Mehrjardi, 2016). However, some studies did not report the SDs used for their 3D-DSM studies (de Carvalho et al., 2014), and many studies utilized legacy soil profiles without specific SD (Adhikari et al., 2013; Malone et al., 2009; Meersmans et al., 2009). For 3D-DSM studies, more attention has been paid to improving the 3D mapping techniques and interpreting the final map products. Although various SDs have been widely used in 3D-DSM, limited research has focused on the contribution of SD in the map accuracy. In addition, a SD with good performance in surface soil might not guarantee a good performance in deep soil. Therefore, the comparison and assessment of performance of different SDs in multiple layers are necessary for investigation.

In addition, the development of 3D-DSM techniques imposed higher standard on SDs. The 3D regression-kriging (RK), with rapid, accurate, and automated mapping procedures and flexibility of using various regression techniques, has become popular and been adopted for global soil grid maps (Hengl et al., 2014). Two integral components of RK are the regression method for trend prediction and the interpolation technique for interpreting spatial structure (Hengl et al., 2015). An accurate estimation of regression coefficients in the calibration model depends heavily on a good spread in feature space, while a good dispersion in geographical space greatly contributes to a reliable interpolation of sample data (Brus and Heuvelink, 2007). Therefore, a sound SD with both good spatial coverage and feature space coverage should be identified for 3D RK and assessed for its reliability for mapping multiple depths.

The main purpose of this paper was to compare six SDs, including: 1) grid sampling (GS), 2) grid random sampling (GRS), 3) simple random sampling (SRS), 4) stratified random sampling (StRS),

5) transect sampling (TS), and 6) conditioned Latin hypercube sampling (cLHS) for calibrating RK models in 3D-DSM.

5.2 Materials and methods

A detailed description of study area, spectra data collection and processing, soil sample collection and analysis, and spectral model (PLSR) could be found in chapter 3. In addition, environmental covariates and the procedure of 3D-DSM (RK-RF) was discussed in chapter 4.

5.2.1 Sampling designs

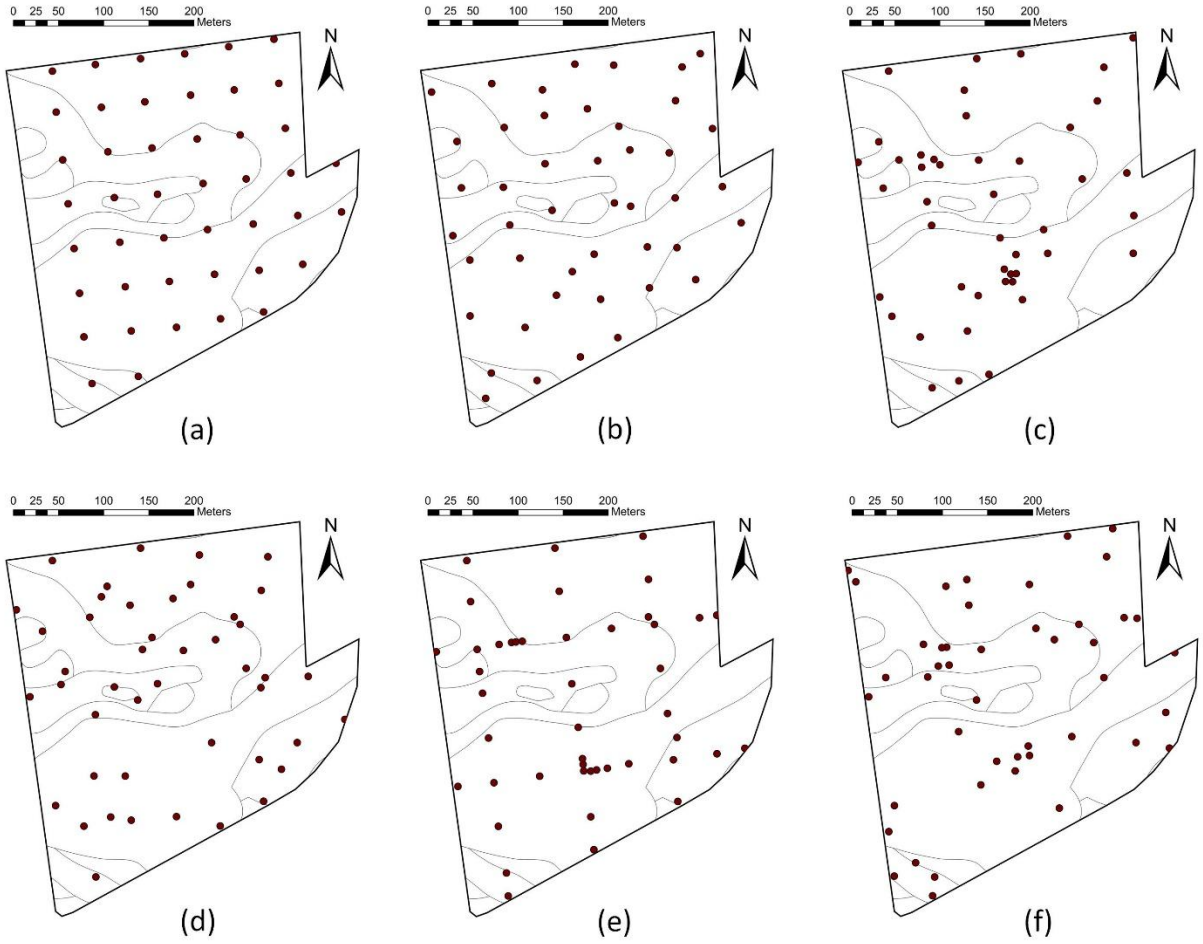

Fig. 5.1. A total of 45 sampling points identified by six different sampling. (a) Grid sampling (GS); (b) Grid random sampling (GRS); (c) Simple random sampling (SRS); (d) Stratified random sampling (StRS); (e) Transect sampling (TS); (f) conditioned Latin hypercube sampling (cLHS). The lines represent soil type boundary following a detailed soil survey done in 1971.

A total of 148 sample locations were identified following each of six different SDs including GS, GRS, SRS, StRS, TS, and cLHS (Fig. 5.1). Every SD consists of a set of 45 sample locations. A

square grid with intervals of 25m was used to create GS in this study. A sample was randomly selected within each grid of the GS to produce GRS. SRS is the most basic and common probability based SD, in which each unit is selected randomly and independently (Webster and Lark, 2013). In StRS, elevation was used to assist in stratification of the field, and then sample locations were proportionally selected in each stratum according to the area of that stratum. Five transects (3 north-south transects and 2 east-west transects) were placed in this field, and sample locations were unequally (nested) selected in each transect.

Latin hypercube sampling (LHS) is a maximally stratified random sampling procedure that achieves a full coverage of multivariate distributions. Furthermore, in order to obtain an approximate LHS from an available dataset, conditioned Latin hypercube sampling (cLHS) was proposed by adding a search algorithm based on heuristic rules and an simulated annealing (SA) process on LHS (Minasny and McBratney, 2006). In this study, based on the available 148 locations, cLHS was conducted to spread 45 sample locations in the ten environmental covariates and further used to compare with other SDs which were mainly optimized in geographical space. The soil properties at the selected 45 sites according to each SD were individually used in the DSM to create sub maps by the 3D RK-RF method which will be further compared with the original maps produced by total 148 sample sites.

5.2.2 Complete spatial randomness (CSR)

Complete spatial randomness test (CSR), also known as spatial Poisson process, examines whether a spatial point pattern in a given area occurs in a completely random fashion (Maimon and Rokach, 2005). In other words, it is to test whether the point pattern is independently and uniformly distributed over an area, rather than interacting with each other. A single factor that influences the test results is the density of points. The nearest neighbor distance distribution function of a stochastic point process calculates the cumulative distribution function G(r) against the distance (r) from certain random points to the nearest other point in the stochastic process (Baddeley et al., 2007). This function was used to form a theoretic CSR which was compared with the real point pattern of this study.

5.2.3 Optimization criteria

S-optimality criterion was used to assess the spatial separation of selected sample locations by calculating the horizontal distances among pairs of these locations (Adamchuk et al., 2011). It

seeks to maximize the harmonic mean distance from each sample location to all the other locations in the SD (SAS, 2008):

$$S_{opt} = \frac{N(N-1)}{2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{1}{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}}} \quad (5.1)$$

where N is the sample size (45 in this study), x and y are the geographical coordinates of the i^{th} and j^{th} locations.

D-optimality was selected to assess the degree of variability of selected dataset by every sampling design. It increases with the greater coverage of variables by selected dataset (Adamchuk et al., 2011). D-optimality was applied on the premise of the linear assumption between soil properties and environmental variables.

$$D_{opt} = |Z'Z| \quad (5.2)$$

$$Z = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_N \end{bmatrix} \quad (5.3)$$

where z_i is the soil property or environmental variable for i^{th} location.

However, one limitation of the D-optimality is that only one variable is taken into account. In DSM, we are more interested to assess the coverage of multiple variables either soil properties of interest or environmental variables in the feature space. Therefore, a modified criterion that can simultaneously assess the multivariate distribution is required.

In order to test the coverage of selected sample points in feature space in this study, Z was extended from single variable to multiple variables:

$$Z = \begin{bmatrix} 1 & z_{11} & z_{21} & z_{21} & \cdots & z_{k1} \\ 1 & z_{12} & z_{22} & z_{32} & \cdots & z_{k2} \\ 1 & z_{13} & z_{23} & z_{33} & \cdots & z_{k3} \\ 1 & z_{14} & z_{24} & z_{34} & \cdots & z_{k4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{1n} & z_{2n} & z_{3n} & \cdots & z_{kn} \end{bmatrix} \quad (5.4)$$

where Z is the environmental variables matrix, k is the number of variables (j=1:10, 10 environmental variables used in this study), n is the number of observations (i=1:133 for the whole dataset, and i=1:45 for each SD), z_{kn} is the value of j^{th} environmental variable in the i^{th} location.

Therefore, the D_{opt} become:

$$D_{opt} = |Z'Z| = \left\| \begin{bmatrix} 1 & \sum_{i=1}^{N} z_1 & \sum_{i=1}^{N} z_2 & \sum_{i=1}^{N} z_3 & \cdots & \sum_{i=1}^{N} z_k \\ \sum_{i=1}^{N} z_1 & \sum_{i=1}^{N} z_1^2 & \sum_{i=1}^{N} z_1 z_2 & \sum_{i=1}^{N} z_1 z_3 & \cdots & \sum_{i=1}^{N} z_1 z_k \\ \sum_{i=1}^{N} z_2 & \sum_{i=1}^{N} z_1 z_2 & \sum_{i=1}^{N} z_2^2 & \sum_{i=1}^{N} z_2 z_3 & \cdots & \sum_{i=1}^{N} z_2 z_k \\ \sum_{i=1}^{N} z_3 & \sum_{i=1}^{N} z_1 z_3 & \sum_{i=1}^{N} z_2 z_3 & \sum_{i=1}^{N} z_3^2 & \cdots & \sum_{i=1}^{N} z_3 z_k \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} z_k & \sum_{i=1}^{N} z_1 z_k & \sum_{i=1}^{N} z_2 z_k & \sum_{i=1}^{N} z_3 z_k & \cdots & \sum_{i=1}^{N} z_k^2 \end{bmatrix} \right\| \quad (5.5)$$

5.2.4 Validation

For every SD, except for the 45 sites which were identified for DSM, the remaining 103 sites were used for validating the accuracy of maps. RMSE values were calculated for 10 depths of every soil property.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5.6)$$

where *n* was the number of samples, y_i was the measured value, and $\hat{y}_i$ was the predicted value.

5.2.5 Data analysis

Sample locations selected by cLHS method were optimized from eight environmental covariates with 10000 iterations by the 'clhs' package (Roudier and Roudier, 2015) in R version 3.2.3 (The R Foundation). The CSR test was implemented by 'spatstat' package in R (Baddeley et al., 2015) version 3.2.3 (The R Foundation). The maps were produced by RK-RF in the 'GSIF' package (Hengl, 2014) in R version 3.2.3 (The R Foundation). The S-optimality and D-optimality were analyzed in MATLAB R2015b (The Mathworks Inc., MA, USA).

5.3 Results and discussions

5.3.1 Spatial and feature space coverage

Based on the CSR results of the whole dataset (Fig. 5.2), the spatial point pattern did not exactly satisfy the CSR test. It appeared that the observed G-curve fitted into the 95% range of the CSR estimated in the shorter distance and longer distance. This indicated that there was no clusters in the shorter distance or under-sampled regions in the longer distance. However, the observed G-

curve in the middle range distances was lower than the theoretical G-curve, indicating that the samples were over-sampled (Bivand et al., 2008).

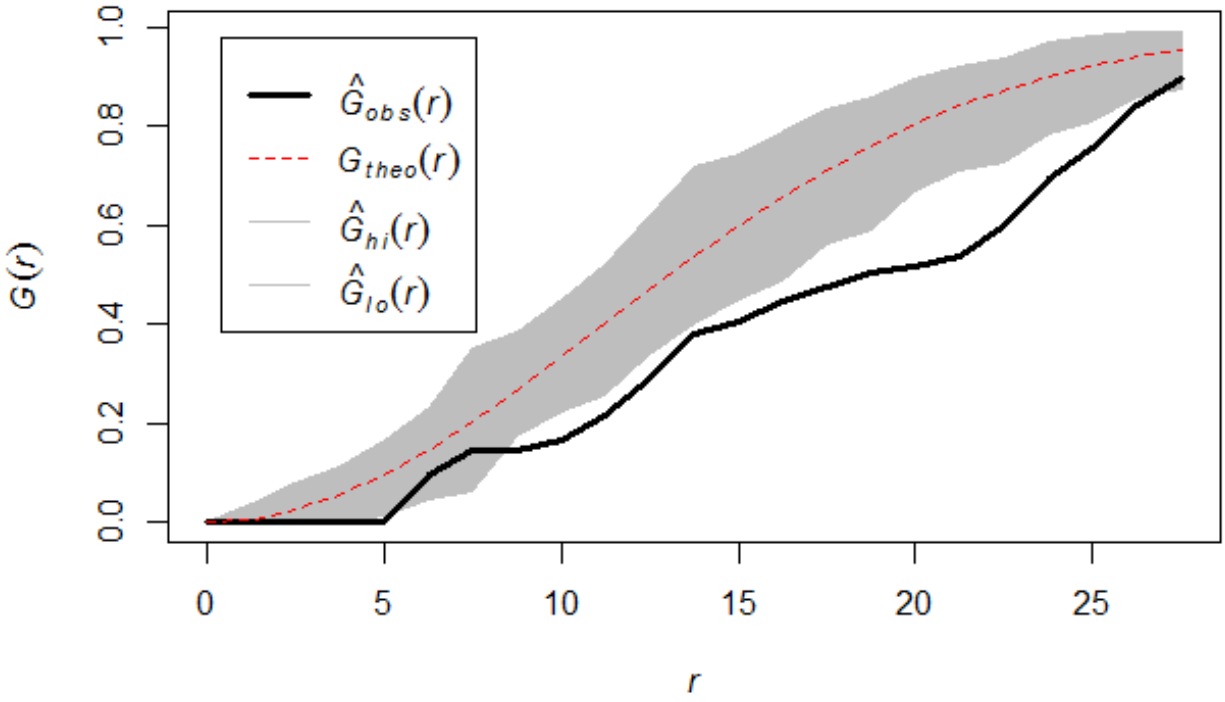

Fig. 5.2. CSR of the whole dataset with 148 sample sites. G_theo (r) is the theoretical complete spatial randomness, with the upper boundary (Gˆ_hi (r)) and the lower boundary (Gˆ_lo (r)) of 95% confidence interval, and Gˆ_obs (r) is the actual pattern of the 148 sample points.

Table 5.1 S-optimality and D-optimality test of six SDs with sample size of 45

	GS	GRS	SRS	StRS	TS	cLHS
S_{opt}	**135.13**	133.54	104.64	123.59	106.93	119.21
$D_{opt\ (\times 10^{23})}$	1.28	5.97	12.6	8.13	0.72	**18.4**

The bold indicates the highest S_{opt} and the D_{opt} among the six SDs: GS- grid sampling, GRS- grid random sampling, SRS- simple random sampling, StRS- stratified random sampling, cLHS- conditioned Latin hypercube sampling.

GS showed the best spatial coverage with the highest S_{opt} of 135 followed by GRS with the S_{opt} of 134 (Table 5.1). This is consistent with the previous study by Adamchuk et al. (2011) that the rectangular grid sampling obtained relatively higher S_{opt}. The S-optimality results aligned well to the visual comparison that the GS obtained the most even coverage and GRS spread the sample points uniformly throughout the study area (Fig. 5.1). The StRS with the S_{opt} of 124 obtained fair spatial coverage with under-sampled regions in the lower part of the area. cLHS with the S_{opt} of

120 obtained moderate spatial coverage with some clusters and more under-sampled area compared to StRS. Adamchuk et al. (2011) also obtained medium S_{opt} for LHS. However, the cLHS (Minasny and McBratney, 2006) did not form clusters and the sample points were well spread in the geographical space. TS with low S_{opt} of 107 showed very poor spatial coverage as all the sample points were arranged along straight lines. SRS with the lowest S_{opt} of 105 formed many clusters and under-sampled area.

D-optimality illustrated the coverage in feature space formed by eight environmental variables in this study. The cLHS with the D_{opt} of 18.4 greatly surpassed all the other SDs, since it was produced by maximally stratifying the variations in the same multivariate distributions. In addition, the StRS that utilized the elevation information in sample site selection also showed good D_{opt} of 8.13. By comparing the GS and GRS, the greater D_{opt} of GRS over GS might be due to the flexibility gained by the randomness procedure. The TS with the D_{opt} of 0.72 showed lowest feature space coverage. However, SRS, surprisingly, obtained very good feature space coverage with D_{opt} of 12.6, while still worse than cLHS. The similar results were obtained by Mulder et al. (2013) that cLHS obtained a good coverage in feature space in comparison with SRS and GS. Minasny and McBratney (2006) also obtained good feature space coverage for cLHS while a biased coverage for SRS by comparing the histogram and boxplot of the distribution.

5.3.2 Sub maps vs. original map

By comparing the sub maps produced by different SDs and the original maps, StRS comparatively performed best in reproducing the spatial pattern of the original map for all the soil properties. SRS in general outperformed other SDs in the top three layers, but overestimated the low values for GWC and underestimated the high values for pH and K in deep soil. GS and TS overestimated the values for BD and pH throughout the profile, and underestimated the values in deep soil for VWC, clay content (Fig. 5.7), K, Mg, and Mn (Figures of these properties not shown in this thesis). GRS and cLHS slightly underestimated the high and low values of soil properties in the bottom left and middle part of the study area throughout the whole soil profile and showed smoothness effect for many soil properties including GWC, BD, SOM, EC, Na, and Zn. In addition, GRS underestimated the values for pH, P (Fig. 5.9), sand content (Fig. 5.5), and Mg and cLHS underestimated the values for VWC and K, respectively. GRS performed well in deep soils for VWC, clay content, K, and Mn, as it effectively exhibited the high values in these soil properties in deep soil. Although cLHS produced smooth effects of highest and lowest values of the sub

maps, it mimicked the general distribution slightly better than others SDs for many soil properties. However, this was a rough and visual comparison between sub maps and the original map, a more robust way by comparing the point to point values was presented by RMSE values (Fig. 5.3).

The lowest standard errors were mainly observed in StRS for VWC, GWC, BD, EC, sand content (Fig. 5.6), clay content (Fig. 5.8), Mn, and Zn and in GS for GWC, BD, SOM, pH, EC, P (Fig. 5.10), K, Na, and Mg (figures of other soil properties not shown in this thesis). GRS resulted in relatively low standard errors for almost all the soil properties throughout the whole study area, though slightly higher than GS and StRS. In addition, other SDs occasionally obtained standard errors as low as GS and StRS. For example, the low standard errors of clay content were obtained in TS and StRS; the low standard errors for Mg were obtained in GS and TS. TS also resulted in a low standard error for pH. GRS, SRS, and cLHS showed clearly lower standard errors for Mn, K, and SOM, respectively. However, SRS, TS, and cLHS obtained slightly worse results, mainly in the deep soil layers at the right edge of the field where the soil cores only reached 40 to 50 cm. In TS, it was visually clear that the area around the transect lines exhibited smaller standard errors than the area far from transect lines. A similar feature in standard error distribution was also observed through the fragmented colors in SRS and cLHS. This feature proved that the uncertainties were mainly determined by the amount of available information provided by the samples. Therefore, a better spatial coverage always led to a smaller standard error across the whole study area. Furthermore, the good coverage in feature space might have partly contributed to the smaller uncertainty in StRS while moderate geographical coverage resulted in many lowest standard errors.

However, the small standard errors or prediction uncertainties did not certainly lead to a high accuracy (Fig. 5.3). The GS with the lowest standard errors in many soil properties did not show any superiority to other SDs in RMSE values when compared with the original map. However, the StRS was substantially better than other SDs almost throughout the whole profile for BD, P, and Zn, and in top soils for VWC, K, and Mg, and in deep soils for GWC, EC, Na, and Mn. In addition, cLHS obtained the lowest RMSE values throughout the whole profile of soil pH, and in the deep soil of VWC. GRS outperformed other SDs in the top soils of GWC, SOM, and Na and deep soils of K and Mg. TS resulted in few low RMSE values for VWC, sand content, clay content, and Na in sub soil, and SRS also showed few low RMSE values for GWC, EC, and Na in sub soil and Mn in top soil. Therefore, StRS with both relatively good spatial coverage and feature space coverage

outperformed other SDs in mimicking the original map distribution. This was slightly different from Minasny and McBratney (2006) who reported that the cLHS better reflected the original distribution compared with SRS and spatial coverage sampling. In that study, the cLHS was selected from a fair large dataset and also represented a good spatial coverage. Additionally, the StRS was not compared in that study. Falk et al. (2011) compared the SRS, GS, StRS, and cLHS with the original distribution and demonstrated that cLHS was the best followed by StRS and SRS with the lowest accuracy.

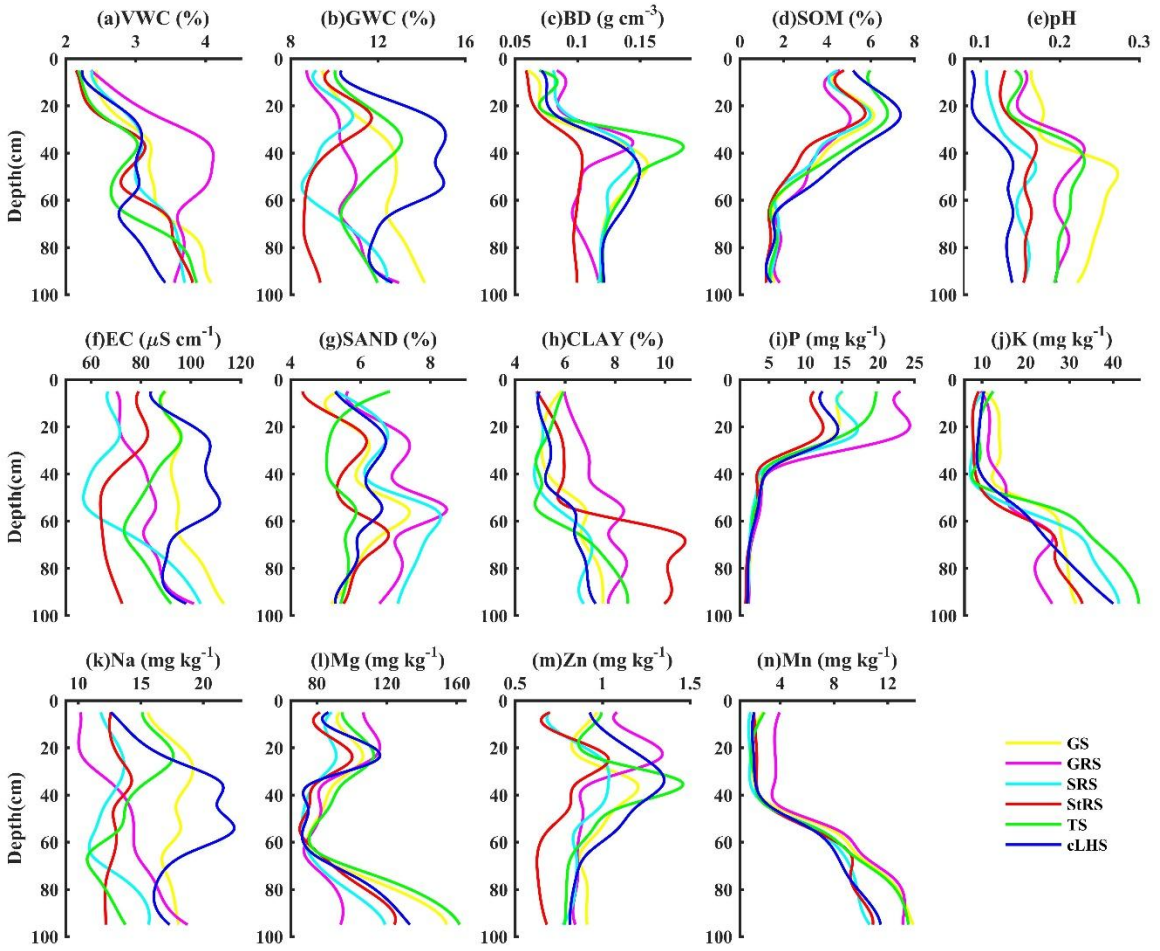

Fig. 5.3. RMSE values between original maps and sub-maps generated following six SDs. GS: grid sampling; GRS: grid random sampling; SRS: simple random sampling; StRS: stratified random sampling; TS: transect sampling; cLHS: conditioned Latin hypercube sampling.

5.3.3 Validation

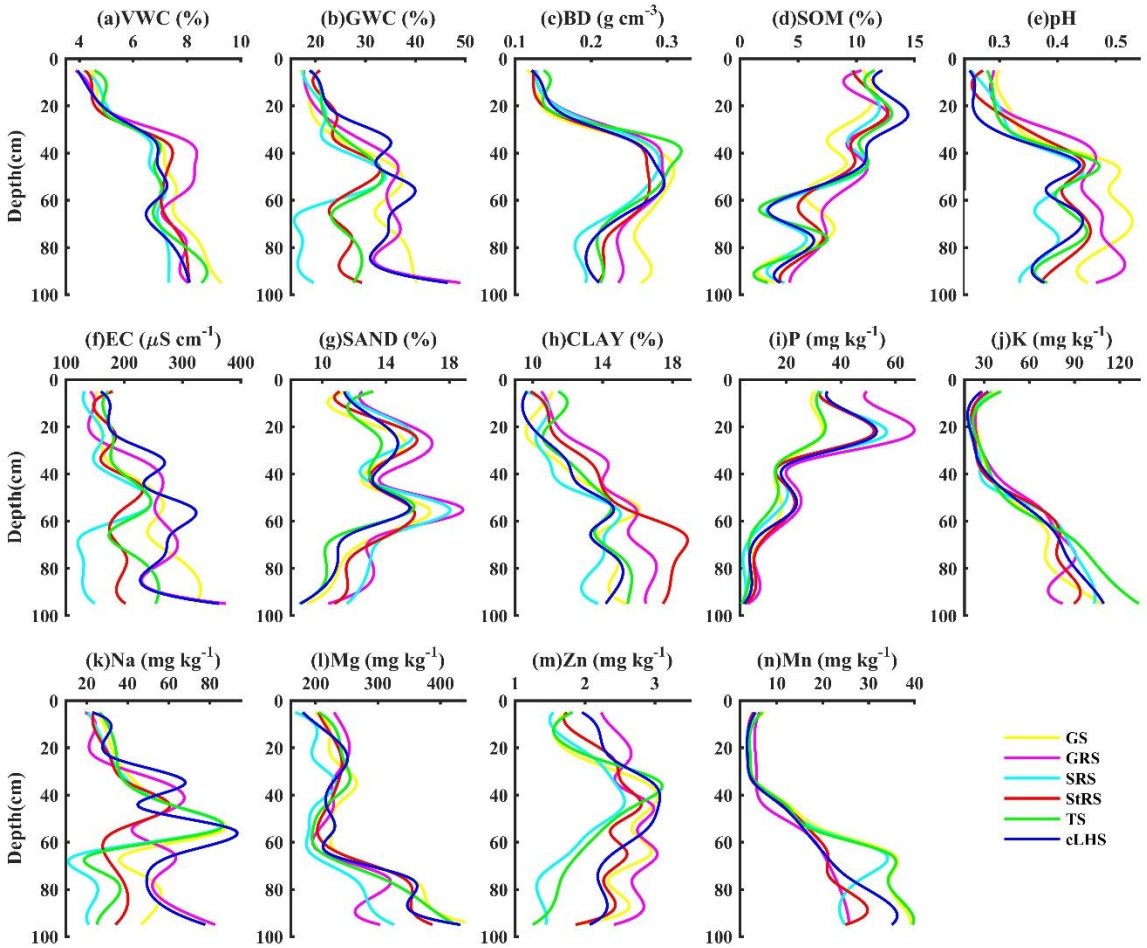

Fig. 5.4. Validation results of six SDs. GS: grid sampling; GRS: grid random sampling; SRS: simple random sampling; StRS: stratified random sampling; TS: transect sampling; cLHS: conditioned Latin hypercube sampling.

The RMSE values of the validation dataset (Fig. 5.4) showed great difference compared to the RMSE values of the original maps (Fig. 5.3) and this proved the necessity of an independent dataset for validation. The RMSE values of StRS in the validation dataset were not as good as its comparison with original maps. Additionally, cLHS achieved moderate results and slightly outperformed other SDs with the smallest RMSE for VWC, pH, clay content, K, and Mn in the top and sub soil layers. However, previous studies reported a high superiority of cLHS to other SDs (Minasny and McBratney, 2007; Worsham et al., 2012). Schmidt et al. (2014) even resulted that cLHS outperformed other model-based SDs e.g. fuzzy k-means sampling and response surface sampling. So far, cLHS has been widely used in soil properties prediction and DSM and achieved high efficiency, such as 3D mapping of SOC and BD from gamma radiometric emission,

geological variables, and topographic attributes (Lacoste et al., 2014); soil salinity mapping (Taghizadeh-Mehrjardi et al., 2014); and SOC prediction from vis-NIR spectroscopy (Kanika et al., 2012). Furthermore, in order to meet the practical accessibility issues by reducing travel time, a cost function has been added as a covariate layer to increase the sampling efficiency (Clifford et al., 2014; Mulder et al., 2013; Roudier et al., 2012). Other constraints and available environmental variables could also be added as covariates, and this greatly increases the efficiency and flexibility. Therefore, cLHS is currently the most widely used and highly recommended sampling design for its advantages of full coverage in feature space and feasibility to add more criteria to achieve specific purpose.

It is noteworthy that SRS which was often reported to have the lowest efficiency and accuracy showed slightly better results in the validation dataset. It achieved better prediction in many horizons of almost all the soil properties, especially in the deep soil for GWC, BD, pH, EC, clay content, P, Na, and Zn, and whole profiles for Mg. The good results obtained by SRS might be due to its good feature space coverage and this effect was more significant in deep soil. However, as the most fundamental SD, SRS has rarely been used in DSM. The randomness of SRS increased its flexibility and contributed to a better result in this study. At the same time, randomness also reduced its robustness for its generality and practical application. Therefore, SRS's feasibility and efficiency in DSM needs further exploration.

In addition, GS achieved smaller RMSE mainly in the top soil, such as sand content and P. But simultaneously it achieved larger RMSE in many soil properties in deep soil compared to other SDs, including VWC, GWC, BD, pH, EC, and Mn. Falk et al. (2011) and Thomas et al. (2012) reported a lower accuracy obtained by SRS compared to GS, but both of these were worse than StRS and cLHS. GS has been widely used in either 2D or 3D DSM for mapping various soil properties, such as soil compaction (Veronesi et al., 2012), soil pH (Vašát et al., 2012), EC (Michot et al., 2013), SOC (Malone et al., 2011), SOM (Poggio et al., 2013), clay content (Huang et al., 2014a), and A horizon thickness. However, GS did not show either superiority to other SDs, or extremely worse results in this study. Similarly, TS did not outperform other SDs, while it obtained comparatively low RMSE in several layers of many soil properties, especially for SOM, sand content, P, Na, and Zn. TS was rarely used alone in DSM, but often applied as a complementary to other SDs, e.g. StRS (Cambule et al., 2014), GS (Samyn et al., 2012), cLHS (van Zijl et al., 2014), purposive sampling (Miller and Schaetzl, 2015) in order to increase the understanding of

pedogenetic processes or form clusters to increase the efficiency. Furthermore, sampling along a topo-sequence is an effective strategy and commonly used in mountainous regions (Liess et al., 2012), and even combined with nested sampling so as to capture the multi-scale variation (Thomas et al., 2012). GRS, in general, showed worse results for many soil properties in multiple layers, including VWC, pH, sand content, clay content, P, Na, and Zn.

Many soil properties, including GWC, BD, pH, EC, clay content, K, Na, and Mn showed relatively uniform RMSE values of six SDs in top soil, while increased and diverse RMSE values in deep soil. This might be partly due to the high variability of these soil properties in deep soil and the complexity and difficulty of soil mapping in deep soil. In addition, the high variability of P in top soil also amplified the RMSE values and spread the values of six SDs. The D-optimality values reflecting the feature space coverage played a vital role in determining the map accuracy over S-optimality (Table 5.1). The cLHS with D_{opt} of 18.6 was higher than other SDs, leading to good accuracy in the validation dataset. SRS with good D_{opt} of 12.6 and extremely low S_{opt} also resulted in good accuracy in the validation dataset. Although StRS with D_{opt} of 8.13 did not surpass previous two SDs, it obtained moderate results compared to GS, GRS, and TS. GS with D_{opt} of 1.28 corresponded to a lower accuracy compared to cLHS, SRS, and StRS. But GRS with both acceptable spatial and feature space coverage did not surpass other SDs, and TS with small S_{opt} and D_{opt} values also showed strengths in specific conditions.

5.4 Conclusions

Six different sampling designs, including grid sampling, grid random sampling, simple random sampling, stratified random sampling, transect sampling, and conditioned Latin hypercube sampling, were compared for their ability to provide reliable input data into three-dimensional digital soil mapping with regression-kriging method. Random forest regression method was used for RK due to its superiority in modeling non-linear soil-landscape relationships and for showing higher prediction accuracy and smaller prediction uncertainty.

Grid sampling displayed the most even geographical space coverage, while conditioned Latin hypercube sampling obtained better coverage in feature space. By comparing the sub maps produced by six different sampling designs and the original map, StRS better reflected the spatial distribution of the original maps, followed by SRS. GS and TS slightly overestimated some properties. While GRS and cLHS underestimated both the high value and low values of the some soil properties. However, despite the smoothing effects, cLHS reproduced the spatial distribution

of original maps better than other SDs. Standard errors of corresponding soil property maps produced smaller prediction uncertainty for GS and StRS followed by GRS. In addition, the closer to the sampling transects or sampling locations, the smaller standard errors were obtained. This showed that the available information provided by sample locations was the major reason for smaller uncertainty. Furthermore, comparatively smaller RMSE values between sub maps and original maps were observed by StRS over other SDs.

An independent validation dataset was necessary for determining the map accuracy. A relatively high accuracy was observed by cLHS in the top and sub soil layers and SRS mainly in the deep soil layers. However, the randomness of SRS might restrict it application and further exploration is required for increasing the robustness of SRS. StRS did not show any superiority to other SDs in the validation results. GS with good results in several soil layers could be further explored for the feasibility in DSM. TS with several good results was reported to be widely used in topo-sequence transects or as complementary to other SDs. D-optimality over S-optimality played an essential role in determining the map accuracy of specific SD. Future work should pay attention to improving the optimization criteria of cLHS and increasing the robustness of other sampling designs.

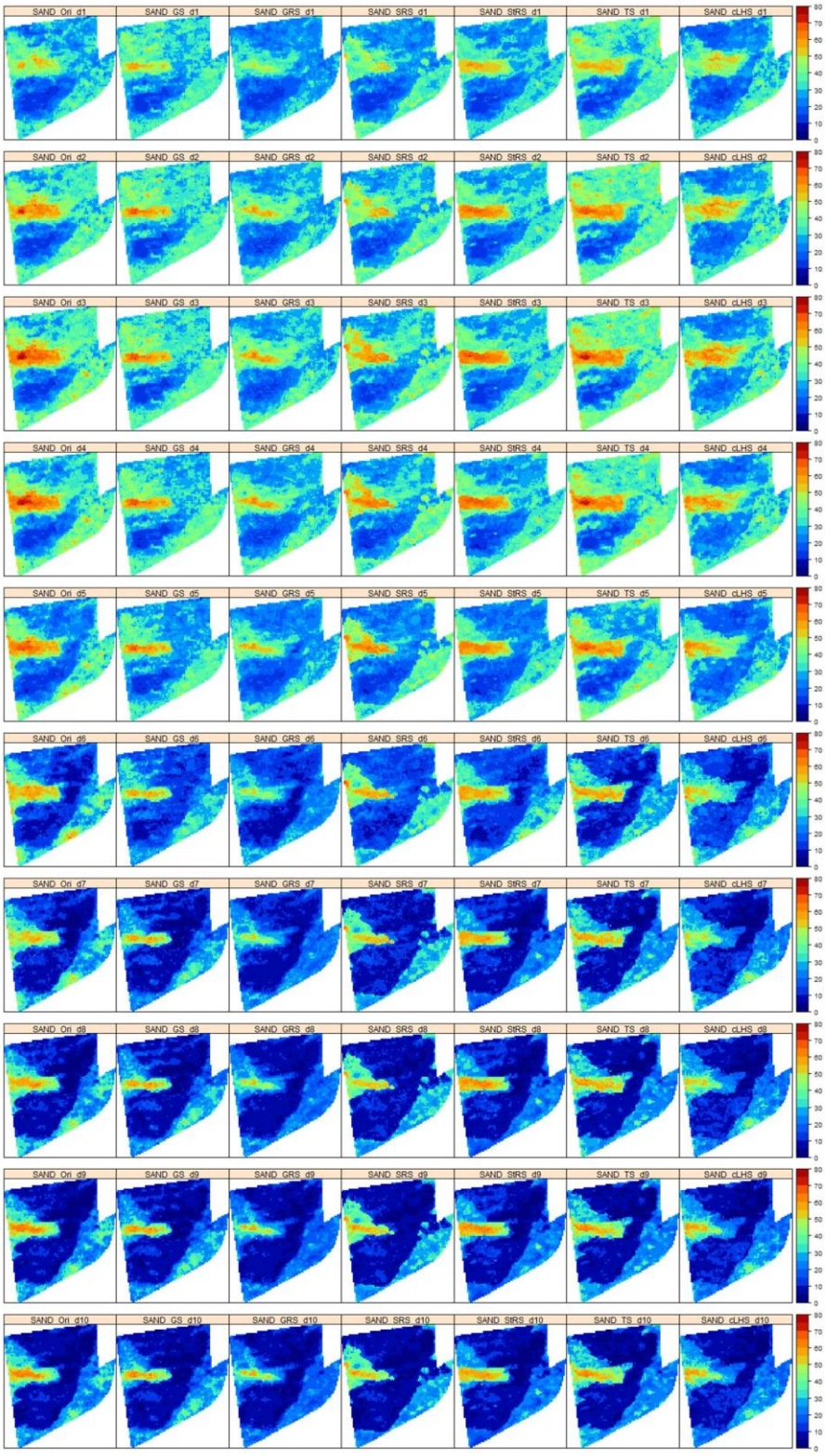

Fig. 5.5. Maps of sand content (SAND) at different depths (d) produced using different sets of data selected by specific sampling designs.

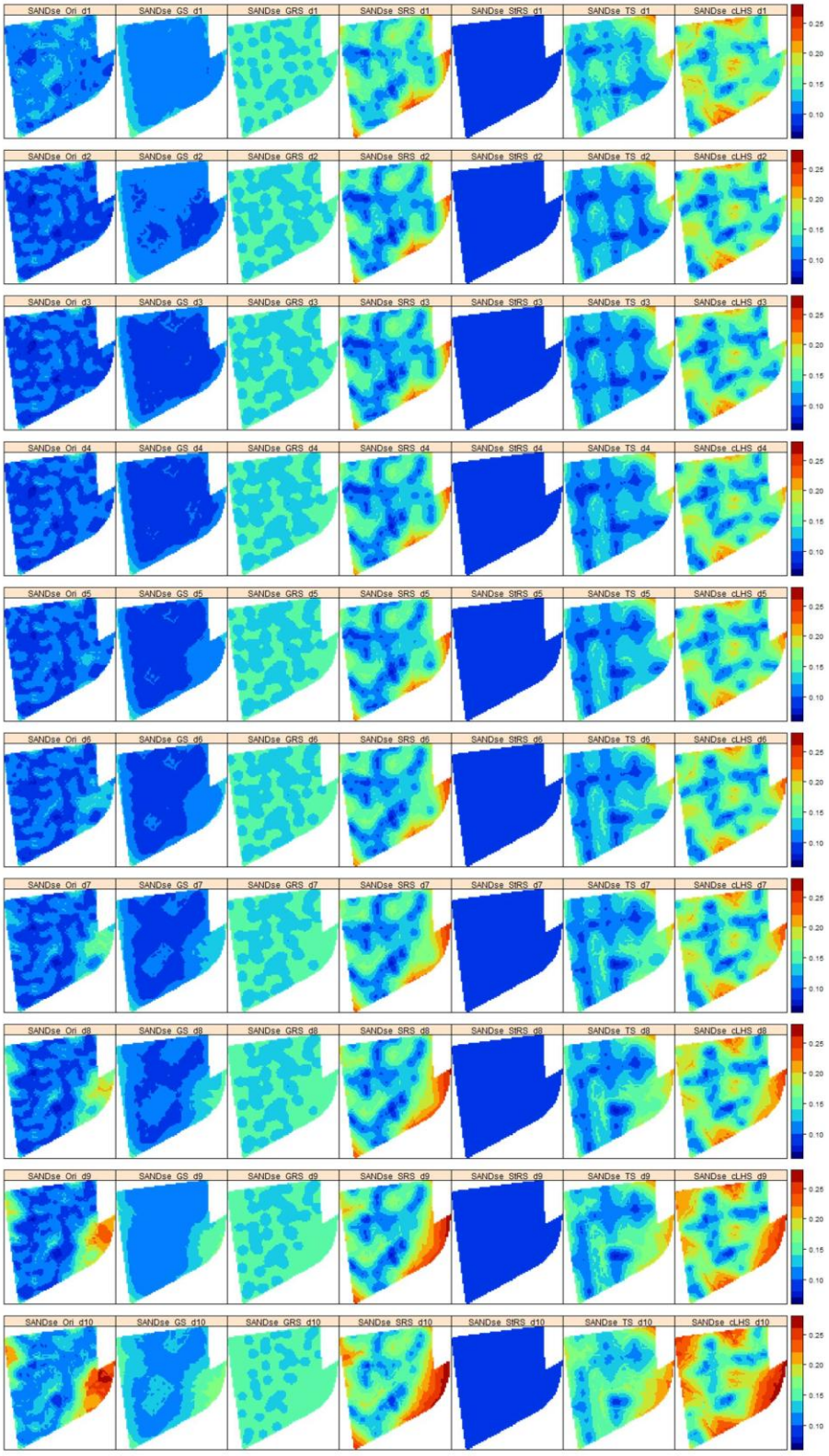

Fig. 5.6. Maps of standard errors of corresponding SAND at different depths (d) produced using different sets of data selected by specific sampling designs.

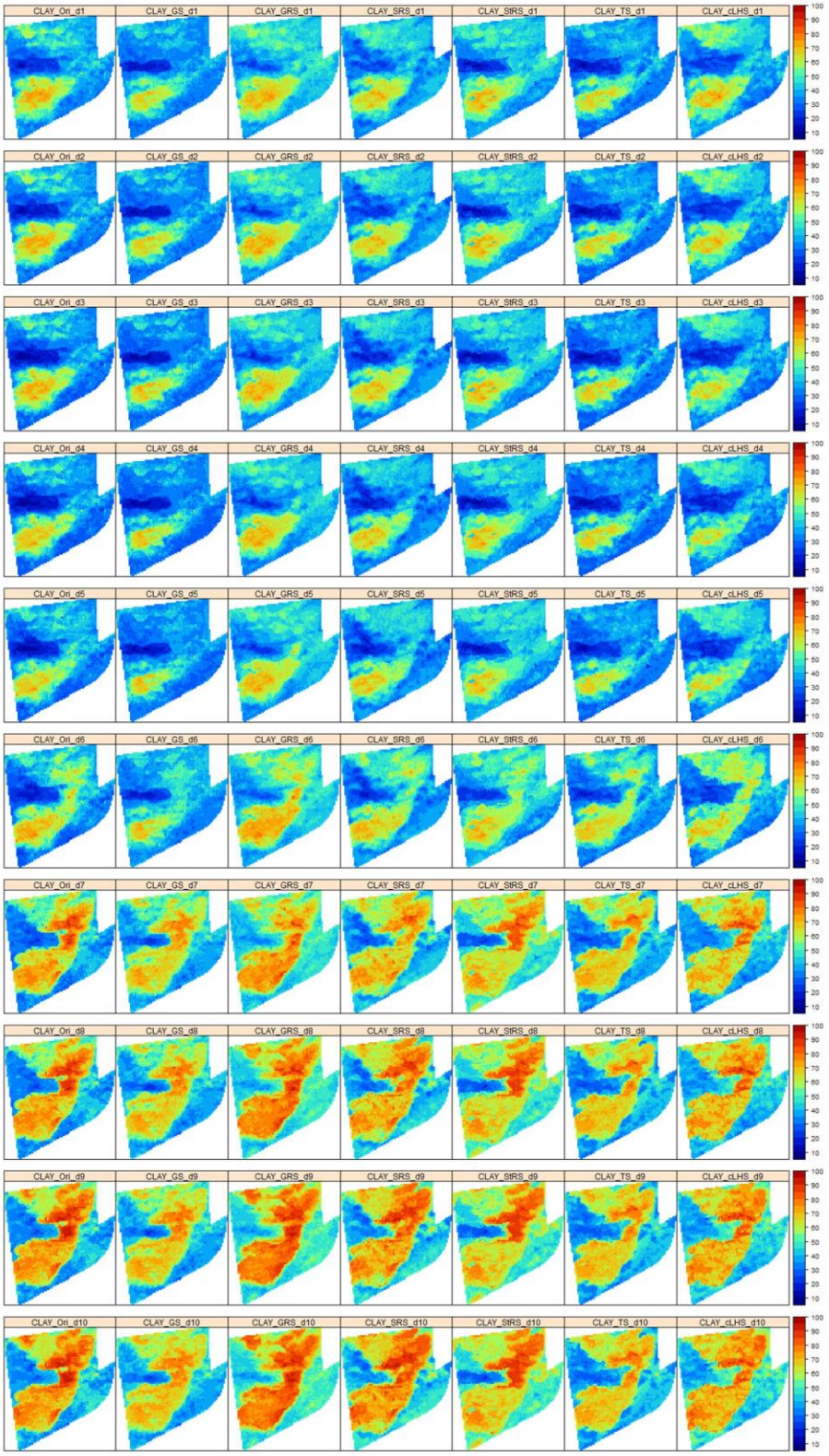

Fig. 5.7. Maps of clay content (CLAY) at different depths (d) produced using different sets of data selected by specific sampling designs.

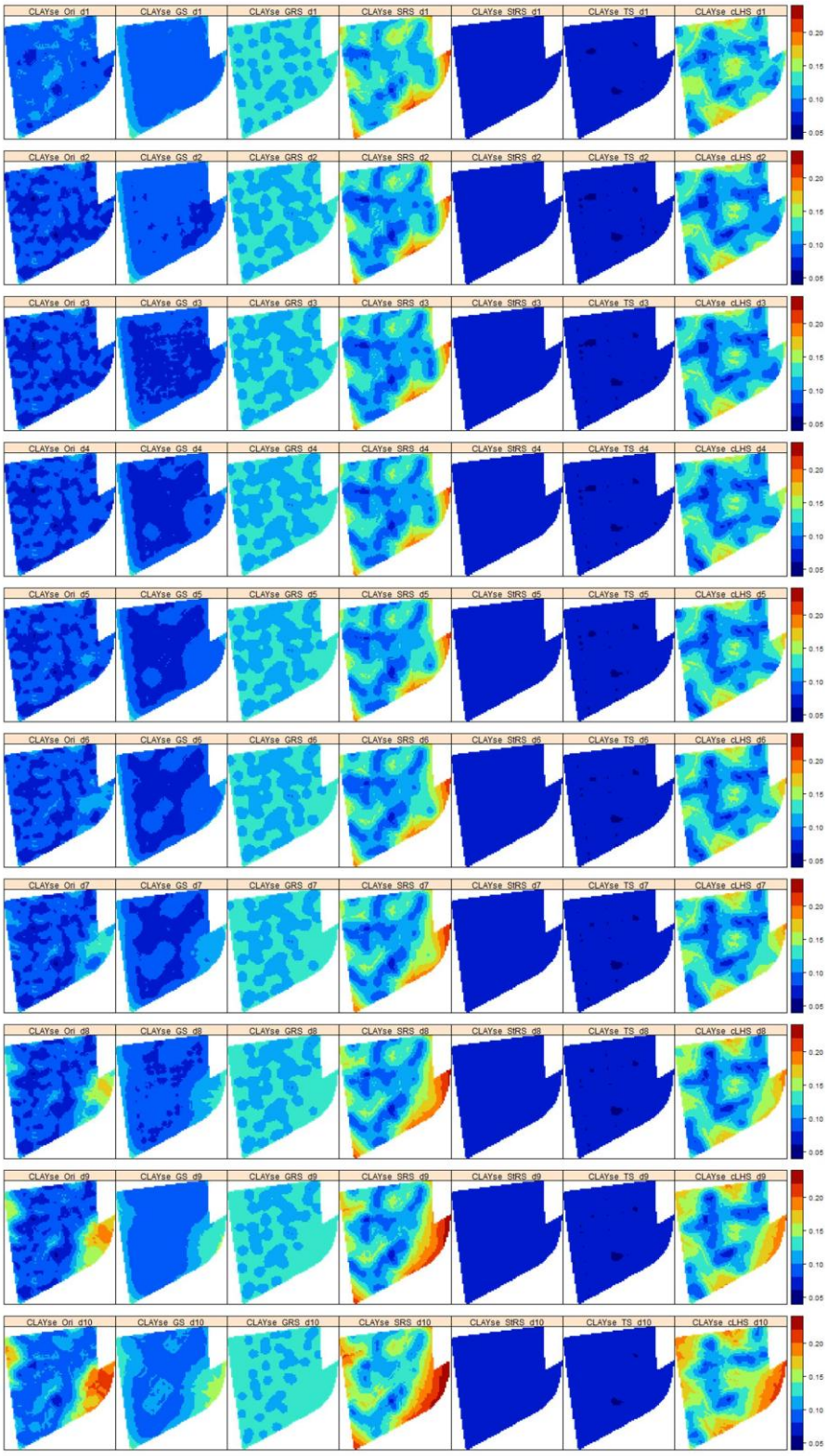

Fig. 5.8. Maps of standard errors of corresponding CLAY at different depths (d) produced using different sets of data selected by specific sampling designs.

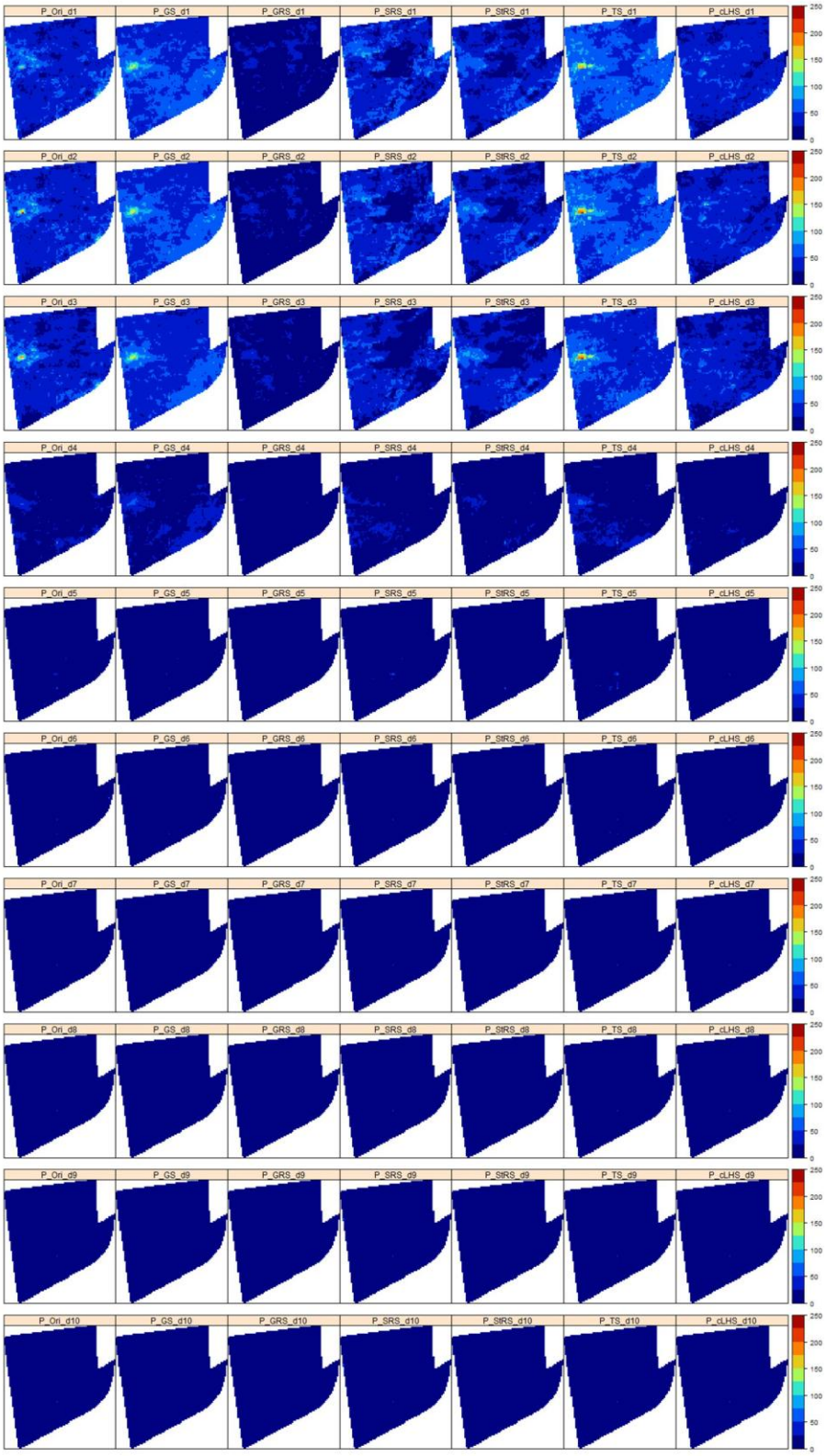

Fig. 5.9. Maps of available phosphorus (P) at different depths (d) produced using different sets of data selected by specific sampling designs.

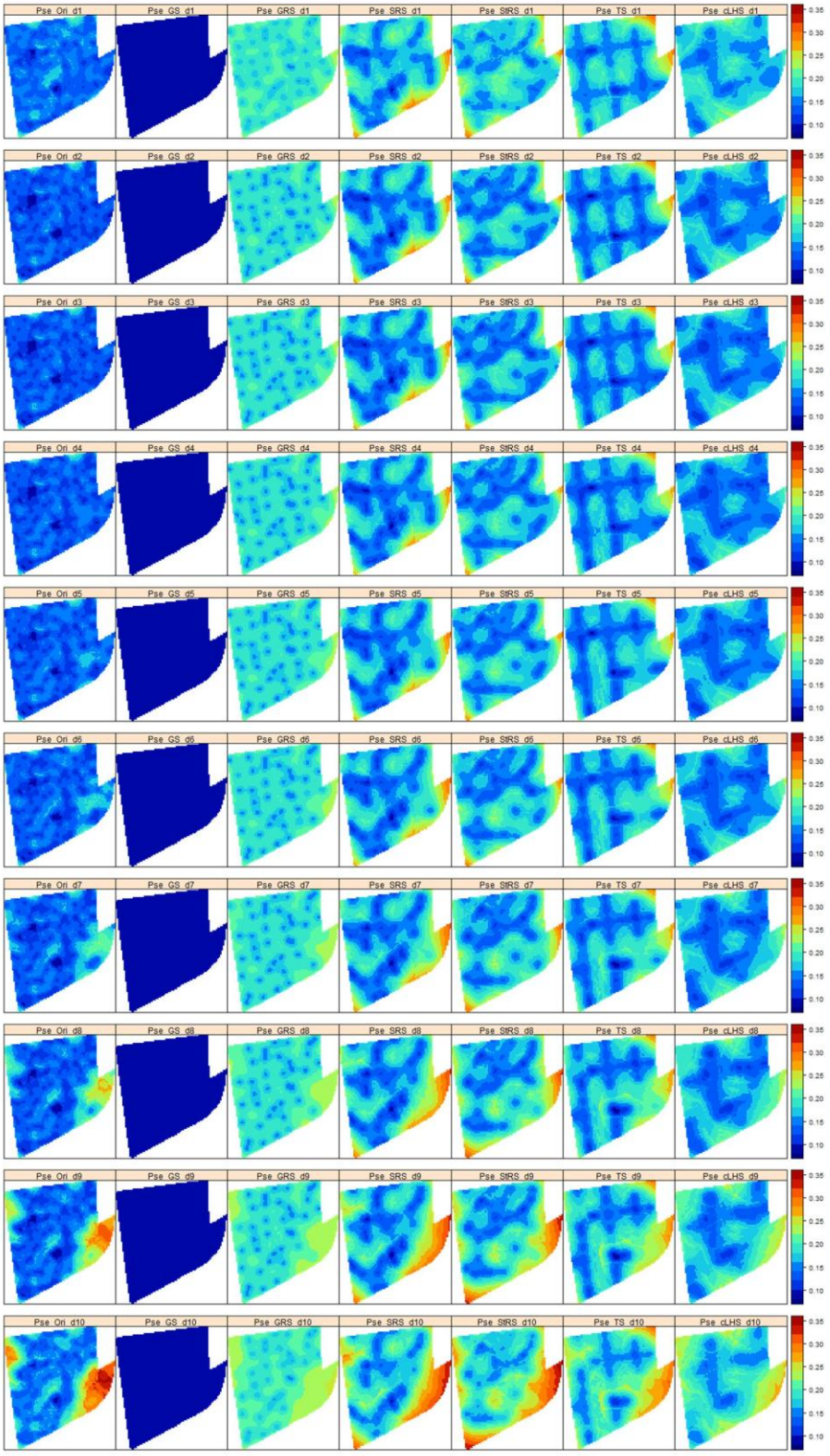

Fig. 5.10. Maps of standard errors of corresponding P at different depths (d) produced using different sets of data selected by specific sampling designs.

CHAPTER 6

General conclusions and future directions

Three-dimensional digital soil maps were prepared for a large number of soil properties using three-dimensional regression-kriging. A set of maps of soil properties as well as corresponding standard errors were displayed as final products. Three regression techniques, including generalized linear model, regression tree, and random forest were adopted and compared to identify the most effective prediction with regression-kriging. In general, RF was regarded as a superior model for RK over GLM and RT due to its capability of interpreting non-linear soil-landscape relationships, fitting weak spatial dependency of regression residuals, and resulting in higher prediction accuracy and smaller prediction uncertainty. In addition, proximal soil sensing techniques (vis-NIR spectra and gamma-ray radiation) were used to densely collect soil information and used as environmental covariates. Soil vis-NIR spectra showed strong and easily recognizable absorption features of SOM and water-related soil properties. Therefore, good predictions were obtained for organic matter and water-related soil properties such as SOM, GWC, BD, EC, and P. Many other soil properties, with direct or indirect relationship with SOM and soil water were reported with fair prediction results with R^2 above 0.5. Therefore, *in-situ* soil vis-NIR spectra was a good technique to easily and exhaustively obtain soil properties for DSM input.

Six different sampling designs, including grid sampling, grid random sampling, simple random sampling, stratified random sampling, transect sampling, and conditioned Latin hypercube sampling were tested and compared for their ability to provide reliable input data into three-dimensional digital soil mapping with regression-kriging method. The most even geographical space coverage was obtained by grid sampling, while conditioned Latin hypercube sampling displayed better coverage in feature space. As a result, StRS with both good spatial and feature space coverage better reflected the spatial distribution of the original maps and resulted in a smaller prediction uncertainty. GS with the most even coverage also presented a smaller prediction uncertainty and this proved that prediction uncertainty was mainly determined by the available information provided into the model. An independent validation dataset was also used to assess the map accuracy. A relatively high accuracy was observed by cLHS mainly in the deep soil layers. SRS showed higher D-optimality in this study, thus resulting in good validation results. It could be further explored for the feasibility and robustness in DSM. TS with several good results was reported to be widely used in topo-sequence transects or as complementation of other SDs.

All the work and results illustrated in this thesis effectively assessed and improved the current techniques in three-dimensional digital soil mapping and contributed to the quantification of the horizontal and the vertical variability of soil properties. The good or fair prediction results obtained from *in-situ* vis-NIR spectra measurement proved the feasibility and efficiency of using proximal soil sensing in 3D-DSM. This is notable because none of previous studies has measured so many soil properties *in-situ* and reached 1m depth. In addition, this was the first time that the latest 3D-DSM method- 3D regression kriging was used at a small-scale for multiple soil properties, and various linear and non-linear regression techniques were simultaneously assessed for the accuracy. This substantially enriched the methodology and practical application of 3D-DSM, provided suggestions and guidance for the further selection and application of techniques for 3D-DSM. Furthermore, as a crucial step of DSM, different sampling designs were also compared and assessed for their contribution to the 3D-DSM and the ability to capture soil variability in multiple layers which have not been discussed in literatures. The results suggested that a small sample selected by stratified random sampling was more efficient to represent the original distribution, while conditioned Latin hypercube was highly recommended for its high flexibility of optimization criteria and accuracy of final maps. Finally, a new profile depth function was proposed to quantify the vertical distribution of soil pH based on the understanding of the pedological and management features of agricultural field, and the generality of this model was tested for a global dataset. This is the first time that the sigmoid model has been developed for quantifying soil spatial variability and could further assist in 3D-DSM of soil pH. To sum up, it can be concluded that through the profile depth function development, proximal soil sensing application, regression techniques and 3D-DSM model selection, and sampling designs comparison, the whole 3D-DSM procedure was assessed and optimized for better quantifying the spatial variability of soil properties.

Several areas were identified for the future work.

1) The vis-NIR spectra were collected in-situ under field moisture and weather condition which might interference the prediction ability on soil properties. Therefore, more analytic and computational work needs to be done for considering and reducing the errors associated with *in-situ* measurements. In addition, the techniques should be improved for timely *in-situ* collection of data.
2) The mapping accuracy in the sub soil and deep soil was relatively low compared to the top soil with available environmental data. Therefore, more work should be paid on

understanding the soil variability and environmental condition in sub and deep soil. In addition, most of the environmental covariates in this study and in the literatures were obtained on the land surface, thus impeding the real prediction and interpretation of sub and deep soil environment, so that more effective and reliable 3D covariates and 3D structure of the models should be explored in the future.

3) As for sampling designs, regression kriging calls for a sampling design with both a good spatial coverage to obtain an accurate estimate of regression coefficients and a good feature space coverage for interpreting the spatial structure and interpolation purpose. Therefore, a sampling design that simultaneously optimizes in geographical space and feature space should be searched. In addition, some sampling design with several good prediction in some soil layers showed unstable prediction accuracy, so that there is a need to optimize and increase the robustness of these sampling designs. Furthermore, cLHS with good feature space coverage and flexibility of adding multiple optimization criteria needs further inquiry. Moreover, various sample sizes could be also compared and determined for an efficient DSM project.
4) The superiority of DSM is not only the high accuracy of final maps, but also the way to quantifying the uncertainty associated. DSM integrates multiple steps, including sampling design and sample collection, spectral model, regression technique, profile depth function, and interpolation method. Each of these steps results in an uncertainty associated with limited knowledge and model prediction. Uncertainty propagation, i.e., how these uncertainties influence the next step of the whole mapping procedure as well as the final products and ultimate uncertainty should be quantified. Therefore, more works to validate the DSM accuracy and uncertainty should be explored.

REFERENCES

2013. Soil Drainage Class. Agriculture and Agri-Food Canada.

Aciego Pietri, J.C., Brookes, P.C., 2008. Relationships between soil pH and microbial properties in a UK arable soil. Soil Biol. Biochem. 40(7), 1856-1861.

Adamchuk, V.I., Rossel, R.A.V., Marx, D.B., Samal, A.K., 2011. Using targeted sampling to process multivariate soil sensing data. Geoderma 163(1), 63-73.

Adamchuk, V.I., Viscarra Rossel, R.A., 2010. Development of On-the-Go Proximal Soil Sensor Systems. In: A.R. Viscarra Rossel, B.A. McBratney, B. Minasny (Eds.), Proximal Soil Sensing. Springer Netherlands, Dordrecht, pp. 15-28.

Adhikari, K., Bou Kheir, R., Minasny, B., Malone, B.P., McBratney, A.B., Greve, M.H., 2012. Continuous depth function mapping af soil pH variability in Denmark, 4th International Congress EUROSOIL 2012.

Adhikari, K., Hartemink, A.E., 2015. Digital Mapping of Topsoil Carbon Content and Changes in the Driftless Area of Wisconsin, USA. Soil Sci. Soc. Am. J. 79(1), 155-164.

Adhikari, K., Hartemink, A.E., Minasny, B., Kheir, R.B., Greve, M.B., Greve, M.H., 2014. Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark. Plos One 9(8), 13.

Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-Resolution 3-D Mapping of Soil Texture in Denmark. Soil Sci. Soc. Am. J. 77(3), 860-876.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle, Selected Papers of Hirotugu Akaike. Springer, pp. 199-213.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., 2014. Digital Mapping of Soil Particle-Size Fractions for Nigeria. Soil Sci. Soc. Am. J. 78(6), 1953-1966.

Al-Asadi, R.A., Mouazen, A.M., 2014. Combining frequency domain reflectometry and visible and near infrared spectroscopy for assessment of soil bulk density. Soil Tillage Res. 135, 60-70.

Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.R., McBratney, A.B., 2014. GlobalSoilMap: Basis of the global spatial soil information system. Taylor & Francis.

Baddeley, A., Bárány, I., Schneider, R., 2007. Spatial point processes and their applications. Stochastic Geometry: Lectures given at the CIME Summer School held in Martina Franca, Italy, September 13–18, 2004, 1-75.

Baddeley, A., Turner, R., Rubak, E., Berthelsen, K.K., Hahn, U., Jalilian, A., van Lieshout, M.-C., Rajala, T., Schuhmacher, D., Waagepetersen, R., 2015. Package 'spatstat'.

Ballabio, C., Fava, F., Rosenmund, A., 2012. A plant ecology approach to digital soil mapping, improving the prediction of soil organic carbon content in alpine grasslands. Geoderma 187, 102-116.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. Applied Spectroscopy 43(5), 772-777.

Batjes, N.H., 2000. Global Soil Profile Data (ISRIC-WISE). ORNL Distributed Active Archive Center.

Ben-Dor, E., Banin, A., 1995. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. Soil Sci. Soc. Am. J. 59(2), 364-372.

Ben-Dor, E., Heller, D., Chudnovsky, A., 2008. A novel method of classifying soil profiles in the field using optical means. Soil Sci. Soc. Am. J. 72(4), 1113-1123.

Ben-Dor, E., Irons, J., Epema, G., 1999. Soil reflettante. Man Remote Sens Remote Sens Earth

Science 3, 111.
Bezdek, J.C., 1981. Pattern-rocognition with fuzzy objective function algorithms. Siam Review 25(3), 442-442.
Bishop, T.F.A., Horta, A., Karunaratne, S.B., 2015. Validation of digital soil maps at different spatial supports. Geoderma 241, 238-249.
Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91(1–2), 27-45.
Bivand, R.S., Pebesma, E.J., Gomez-Rubio, V., 2008. Applied spatial data analysis with R. Springer.
Bot, A., Benites, J., 2005. The importance of soil organic matter: key to drought-resistant soil and sustained food production. Food & Agriculture Org.
Breiman, L., 2001. Random forests. Machine learning 45(1), 5-32.
Brodsky, L., Vasat, R., Klement, A., Zadorova, T., Jaksik, O., 2013. Uncertainty propagation in VNIR reflectance spectroscopy soil organic carbon mapping. Geoderma 199, 54-63.
Bromley, P., 1995. the Effect of Elevation Gain on Soil.
Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132(3), 273-290.
Brown, R.A., McDaniel, P., Gessler, P.E., 2012. Terrain Attribute Modeling of Volcanic Ash Distributions in Northern Idaho. Soil Sci. Soc. Am. J. 76(1), 179-187.
Bruand, A., Hartmann, C., Lesturgez, G., 2005. Physical properties of tropical sandy soils: A large range of behaviours, Management of Tropical Sandy Soils for Sustainable Agriculture. A holistic approach for sustainable development of problem soils in the tropics.
Brungard, C.W., Boettinger, J.L., 2010. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. In: J. Boettinger, D. Howell, A. Moore, A. Hartemink, S. Kienast-Brown (Eds.), Digital Soil Mapping. Progress in Soil Science. Springer Netherlands, pp. 67-75.
Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239, 68-83.
Brus, D.J., de Gruijter, J.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.J., de Gruijter, J.J., van Groenigen, J.W., 2006. Chapter 14 Designing Spatial Coverage Samples Using the k-means Clustering Algorithm. In: A.B.M. P. Lagacherie, M. Voltz (Eds.), Developments in Soil Science. Elsevier, pp. 183-192.
Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138(1-2), 86-95.
Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62(3), 394-407.
Brus, D.J., Noij, I., 2008. Designing sampling schemes for effect monitoring of nutrient leaching from agricultural soils. European Journal of Soil Science 59(2), 292-303.
Brus, D.J., Yang, R.-M., Zhang, G.-L., 2016. Three-dimensional geostatistical modeling of soil organic carbon: A case study in the Qilian Mountains, China. Catena 141, 46-55.
Bryk, M., 2016. Macrostructure of diagnostic B horizons relative to underlying BC and C horizons in Podzols, Luvisol, Cambisol, and Arenosol evaluated by image analysis. Geoderma 263, 86-103.
Bui, E.N., Simon, D., Schoknecht, N., Payne, A., 2006. Chapter 15 Adequate Prior Sampling is

Everything: Lessons from the Ord River Basin, Australia. In: A.B.M. P. Lagacherie, M. Voltz (Eds.), Developments in Soil Science. Elsevier, pp. 193-608.
Burrough, P.A., van Gaans, P.F.M., MacMillan, R.A., 2000. High-resolution landform classification using fuzzy k-means. Fuzzy Sets and Systems 113(1), 37-52.
Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013. A methodology for digital soil mapping in poorly-accessible areas. Geoderma 192, 341-353.
Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., Smaling, E.M.A., 2014. Soil organic carbon stocks in the Limpopo National Park, Mozambique: Amount, spatial distribution and uncertainty. Geoderma 213, 46-56.
Chang, C.-W., Laird, D.A., 2002. Near-infrared reflectance spectroscopic analysis of soil C and N. Soil Science 167(2), 110-116.
Chang, C.-W., Laird, D.A., Hurburgh Jr, C.R., 2005. Influence of soil moisture on near-infrared reflectance spectroscopic measurement of soil properties. Soil Science 170(4), 244-255.
Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65(2), 480-490.
Chaplot, V., Lorentz, S., Podwojewski, P., Jewitt, G., 2010. Digital mapping of A-horizon thickness using the correlation between various soil properties and soil apparent electrical resistivity. Geoderma 157(3-4), 154-164.
Chapman, L., Putnam, D., 1984. The Physiography of Southern Ontario (Ontario Geological Survey, Special Volume 2). Toronto: Ontario Ministry of Natural Resources, 2715.
Chapron, M., 2011. Classification of soil and vegetation by fuzzy K-means classification and particle swarm optimization, International Conference on Swarm Intelligence, France, pp. 1-7.
Chatfield, C., 1995. Model Uncertainty, Data Mining and Statistical-Inference. Journal of the Royal Statistical Society Series a-Statistics in Society 158, 419-466.
Chen, T., Niu, R.-q., Li, P.-x., Zhang, L.-p., Du, B., 2011. Regional soil erosion risk mapping using RUSLE, GIS, and remote sensing: a case study in Miyun Watershed, North China. Environmental Earth Sciences 63(3), 533-541.
Chi, G.Y., Chen, X., Shi, Y., Zheng, T.H., 2010. Forms and profile distribution of soil Fe in the Sanjiang Plain of Northeast China as affected by land uses. Journal of Soils and Sediments 10(4), 787-795.
Clark, R.N., 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy. Manual of remote sensing 3, 3-58.
Clifford, D., Payne, J.E., Pringle, M.J., Searle, R., Butler, N., 2014. Pragmatic soil survey design using flexible Latin hypercube sampling. Comput. Geosci. 67, 62-68.
Cochran, W.G., 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. Annals of Mathematical Statistics 17(2), 164-177.
Corwin, D.L., Lesch, S.M., 2005. Characterizing soil spatial variability with apparent soil electrical conductivity: I. Survey protocols. Computers and Electronics in Agriculture 46(1–3), 103-133.
Corwin, D.L., Lesch, S.M., Segal, E., Skaggs, T.H., Bradford, S.A., 2010. Comparison of Sampling Strategies for Characterizing Spatial Variability with Apparent Soil Electrical Conductivity Directed Soil Sampling. J. Environ. Eng. Geophys. 15(3), 147-162.
Daniel, K.W., Tripathi, X.K., Honda, K., 2003. Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). AUST.

J. SOIL RES. 41(1), 47-59.
de Brogniez, D., Ballabio, C., Stevens, A., Jones, R.J.A., Montanarella, L., van Wesemael, B., 2015. A map of the topsoil organic carbon content of Europe generated by a generalized additive model. European Journal of Soil Science 66(1), 121-134.
de Carvalho, W., Lagacherie, P., Chagas, C.D., Calderano, B., Bhering, S.B., 2014. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma 232, 479-486.
de Gruijter, J.J., Marsman, B.A., 1985. Transect sampling for reliable information on mapping units. Soil spatial variability : proceedings of a workshop of the ISSS and the SSSA, Las Vegas, USA, 30 November - 1 December 1984, 150-165.
de Gruijter, J.J., McBratney, A.B., Taylor, J., 2010. Sampling for High-Resolution Soil Mapping. Proximal Soil Sensing. Springer, Dordrecht.
de Gruijter, J.J., ter Braak, C.J.F., 1990. Model-free estimation from spatial samples: A reappraisal of classical sampling theory. Mathematical Geology 22(4), 407-415.
de Lucena, I.C., Amorim, R.S.S., Lobo, F.D., Baldoni, R.N., Matos, D.M.D., 2014. Spatial heterogeneity of soils of the Cerrado-Pantanal ecotone. Revista Ciencia Agronomica 45(4), 673-682.
de Zorzi, P., Barbizzi, S., Belli, M., Fajgelj, A., Jacimovic, R., Jeran, Z., Sansone, U., van der Perk, M., 2008. A soil sampling reference site: The challenge in defining reference material for sampling. Applied Radiation and Isotopes 66(11), 1588-1591.
Decker, K.L.M., Boerner, R.E.J., 2003. Elevation and vegetation influences on soil properties in Chilean Nothofagus forests. Revista chilena de historia natural 76, 371-381.
Dieterle, F., 2003. Multianalyte Quantifications by Means of Integration of Artificial Neural Networks, Genetic Algorithms and Chemometrics for Time-Resolved Analytical Data.
Diggle, P., Ribeiro, P.J., 2007. Model-based Geostatistics. Springer Science & Business Media.
Diggle, P.J., Tawn, J., Moyeed, R., 1998. Model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics) 47(3), 299-350.
Duffera, M., White, J.G., Weisz, R., 2007. Spatial variability of Southeastern US Coastal Plain soil physical properties: Implications for site-specific management. Geoderma 137(3-4), 327-339.
Evans, D.M., Hartemink, A.E., 2014. Digital soil mapping of a red clay subsoil covered by loess. Geoderma 230, 296-304.
Falk, M.G., Denham, R.J., Mengersen, K.L., 2011. Spatially stratified sampling using auxiliary information for geostatistical mapping. Environ. Ecol. Stat. 18(1), 93-108.
Fernández, F.G., Hoeft, R.G., 2009. Managing soil pH and crop nutrients. Illinois agronomy handbook, 91-112.
Fidencio, P.H., Poppi, R.J., de Andrade, J.C., 2002. Determination of organic matter in soils using radial basis function networks and near infrared spectroscopy. Analytica Chimica Acta 453(1), 125-134.
Fitzgerald, G.J., 2010. Response Surface Sampling of Remotely Sensed Imagery for Precision Agriculture. In: R.A. Viscarra Rossel, A.B. McBratney, B. Minasny (Eds.), Proximal Soil Sensing. Progress in Soil Science. Springer Netherlands, pp. 121-129.
Fitzgerald, G.J., Lesch, S.M., Barnes, E.M., Luckett, W.E., 2006. Directed sampling using remote sensing with a response surface sampling design for site-specific agriculture. Computers and Electronics in Agriculture 53(2), 98-112.
Fritsch, E., Herbillon, A.J., Do Nascimento, N.R., Grimaldi, M., Melfi, A.J., 2007. From Plinthic

Acrisols to Plinthosols and Gleysols: iron and groundwater dynamics in the tertiary sediments of the upper Amazon basin. European Journal of Soil Science 58(5), 989-1006.
Fystro, G., 2002. The prediction of C and N content and their potential mineralisation in heterogeneous soil samples using Vis–NIR spectroscopy and comparative methods. Plant Soil 246(2), 139-149.
Gee, G.W., Bauder, J.W., Klute, A., 1986. Particle-size analysis. Methods of soil analysis. Part 1. Physical and mineralogical methods, 383-411.
Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Analytica Chimica Acta 185(0), 1-17.
Gholizade, A., Soom, M.A.M., Saberioon, M.M., BorůvkaP, L., 2013. Visible and near infrared reflectance spectroscopy to determine chemical properties of paddy soils. Journal of Food, Agriculture and Environment 11(2), 859-866.
Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis. Geoderma 146(1), 102-113.
Grisso, R.D., Alley, M.M., Holshouser, D.L., Thomason, W.E., 2009. Precision Farming Tools. Soil Electrical Conductivity.
Grunwald, S., Thompson, J.A., Minasny, B., Boettinger, J.L., 2012. Digital Soil Mapping in a changing world, Digital Soil Assessments and Beyond. CRC Press, pp. 301-305.
Hartigan, J.A., 1975. Clustering Algorithms. John Wiley \& Sons, Inc.
Hazelton, P.A., Murphy, B.W., 2007. Interpreting soil test results what do all the numbers mean?
Hengl, T., 2014. GSIF: Global Soil Information Facilities. R Package Version 0.5-0.
Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km ? Global Soil Information Based on Automated Mapping. PLoS ONE 9(8), e105992.
Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120(1), 75-93.
Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS ONE 10(6), e0125814.
Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007. About regression-kriging: From equations to case studies. Comput. Geosci. 33(10), 1301-1315.
Hengl, T., Rossiter, D.G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Soil Res. 41(8), 1403-1422.
Huang, J., Lark, R.M., Robinson, D.A., Lebron, I., Keith, A.M., Rawlins, B., Tye, A., Kuras, O., Raines, M., Triantafilis, J., 2014a. Scope to predict soil properties at within-field scale from small samples using proximally sensed gamma-ray spectrometer and EM induction data. Geoderma 232, 69-80.
Huang, J., Subasinghe, R., Triantafilis, J., 2014b. Mapping Particle-Size Fractions as a Composition Using Additive Log-Ratio Transformation and Ancillary Data. Soil Sci. Soc. Am. J. 78(6), 1967-1976.
Huang, J.Y., Barrett-Lennard, E.G., Kilminster, T., Sinnott, A., Triantafilis, J., 2015. An Error Budget for Mapping Field-Scale Soil Salinity at Various Depths using Different Sources of Ancillary Data. Soil Sci. Soc. Am. J. 79(6), 1717-1728.

Islam, K., Singh, B., McBratney, A., 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. Soil Res. 41(6), 1101-1114.
Jafari, A., Finke, P.A., Van de Wauw, J., Ayoubi, S., Khademi, H., 2012. Spatial prediction of USDA- great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. European Journal of Soil Science 63(2), 284-298.
Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw Hill, New York.
Ji, W.J., Shi, Z., Huang, J.Y., Li, S., 2014. In Situ Measurement of Some Soil Properties in Paddy Soil Using Visible and Near-Infrared Spectroscopy. Plos One 9(8), 11.
Jonard, F., Mahmoudzadeh, M., Roisin, C., Weihermuller, L., Andre, F., Minet, J., Vereecken, H., Lambot, S., 2013. Characterization of tillage effects on the spatial variation of soil properties using ground-penetrating radar and electromagnetic induction. Geoderma 207, 310-322.
Joshi, G., Negi, G.C.S., 2015. Physico-chemical properties along soil profiles of two dominant forest types in Western Himalaya. Current Science 109(4), 798-803.
Kahlert, H., Steinhardt, T., Behnert, J., Scholz, F., 2004. A new calibration free pH-probe for in situ measurements of soil pH. Electroanalysis 16(24), 2058-2064.
Kanika, S., Budiman, M., Alex, B.M., Michael, G.S., Fatemeh, N., 2012. Sampling for field measurement of soil carbon using Vis-NIR spectroscopy, Digital Soil Assessments and Beyond. CRC Press, pp. 415-420.
Karunaratne, S.B., Bishop, T.F.A., Baldock, J.A., Odeh, I.O.A., 2014. Catchment scale mapping of measureable soil organic carbon fractions. Geoderma 219, 14-23.
Kawaguchi, K., Kyuma, K., 1974. Paddy soils in tropical Asia. Their Material Nature and.
Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands. Geoderma 241, 313-329.
Kempen, B., Brus, D.J., Stoorvogel, J.J., 2011. Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. Geoderma 162(1-2), 107-123.
Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11(1), 137-148.
Kerry, R., Goovaerts, P., Rawlins, B.G., Marchant, B.P., 2012. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. Geoderma 170, 347-358.
Kerry, R., Oliver, M.A., 2004. Average variograms to guide soil sampling. International Journal of Applied Earth Observation and Geoinformation 5(4), 307-325.
Kerry, R., Oliver, M.A., 2007. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. Geoderma 140(4), 383-396.
Kidd, D., Malone, B., McBratney, A., Minasny, B., Webb, M., 2015a. Operational sampling challenges to digital soil mapping in Tasmania, Australia. Geoderma Regional 4(0), 1-10.
Kidd, D., Webb, M., Malone, B., Minasny, B., McBratney, A., 2015b. Eighty-metre resolution 3D soil-attribute maps for Tasmania, Australia. Soil Res. 53(8), 932-955.
Kidd, D.B., Webb, M.A., Grose, C.J., Moreton, R.M., Malone, B.P., McBratney, A.B., Minasny, B., Viscarra-Rossel, R.A., Cotching, W.E., Sparrow, L.A., Smith, R., 2012. Digital soil assessment, Digital Soil Assessments and Beyond. CRC Press, pp. 3-8.
Knotters, M., Brus, D.J., 2013. Purposive versus random sampling for map validation: a case study

on ecotope maps of floodplains in the Netherlands. Ecohydrology 6(3), 425-434.
Kosmelj, K., Cedilnik, A., Kalan, P., 2001. Comparison of a two-stage sampling design and its composite sample alternative: An application to soil studies. Environ. Ecol. Stat. 8(2), 109-119.
Koszinski, S., Miller, B.A., Hierold, W., Haelbich, H., Sommer, M., 2015. Spatial Modeling of Organic Carbon in Degraded Peatland Soils of Northeast Germany. Soil Sci. Soc. Am. J. 79(5), 1496-1508.
Krueger, J., Böttcher, J., Schmunk, C., Bachmann, J., 2016. Soil water repellency and chemical soil properties in a beech forest soil — Spatial variability and interrelations. Geoderma 271, 50-62.
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.
Lake, J.V., Bock, G., Goode, J., 1997. Precision agriculture : spatial and temporal variability of environmental quality. Wiley, Chichester; New York.
Lal, R., Shukla, M.K., 2004. Principles of soil physics. CRC Press.
Lambe, T., Whitman, R., 1969. Description of an Assemblage of Particles, Soil Mechanics. John Wiley & Sons.
Lame, F.P.J., Defize, P.R., 1993. Sampling of contaminated soil - Sampling error in relation to sample-size and segregation. Environmental Science & Technology 27(10), 2035-2044.
Lark, R.M., 2002. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma 105(1–2), 49-80.
Laslett, G.M., Heuvelink, G.M., Cressie, N., Urquhart, N.S., Webster, R., McBratney, A.B., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil - Discussion. Geoderma 80(1-2), 45-54.
Laycock, P.J., Lopez-Fidalgo, J., 2007. Design of experiments for extreme value distributions. mODa 8 - Advances in Model-Oriented Design and Analysis.
Lesch, S.M., 2005. Sensor-directed response surface sampling designs for characterizing spatial variation in soil properties. Computers and Electronics in Agriculture 46(1-3), 153-179.
Lesch, S.M., D.J., S., 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 Res. 31.
Lesch, S.M., Rhoades, J.D., Corwin, D.L., 2000. ESAP-95 Version 2.01R User Manual and Tutorial Guide.
Levi, M.R., Rasmussen, C., 2014. Covariate selection with iterative principal component analysis for predicting physical soil properties. Geoderma 219, 46-57.
Li, C.F., Cao, C.G., Wang, J.P., Zhan, M., Yuan, W.L., Ahmad, S., 2009. Nitrous Oxide Emissions from Wetland Rice-Duck Cultivation Systems in Southern China. Archives of Environmental Contamination and Toxicology 56(1), 21-29.
Li, H.Y., Shi, Z., Webster, R., Triantafilis, J., 2013. Mapping the three-dimensional variation of soil salinity in a rice-paddy soil. Geoderma 195, 31-41.
Li, H.Y., Webster, R., Shi, Z., 2015a. Mapping soil salinity in the Yangtze delta: REML and universal kriging (E-BLUP) revisited. Geoderma 237, 71-77.
Li, S., Shi, Z., Chen, S.C., Ji, W.J., Zhou, L.Q., Yu, W., Webster, R., 2015b. In Situ Measurements of Organic Carbon in Soil Profiles Using vis-NIR Spectroscopy on the Qinghai-Tibet Plateau. Environmental Science & Technology 49(8), 4980-4987.

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.
Liess, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture Comparison of regression tree and Random Forest models. Geoderma 170, 70-79.
Lillesand, T.M., Kiefer, R.W., 1994. Remote sensing and image interpretation. Wiley & Sons, New York.
Lin, C.Y., Li, P.Z., Cheng, H.G., Ouyang, W., 2015. Vertical Distribution of Lead and Mercury in the Wetland Argialbolls of the Sanjiang Plain in Northeastern China. Plos One 10(4), 10.
Liu, F., Rossiter, D.G., Song, X.D., Zhang, G.L., Yang, R.M., Zhao, Y.G., Li, D.C., Ju, B., 2016. A similarity-based method for three-dimensional prediction of soil organic matter concentration. Geoderma 263, 254-263.
Liu, F., Zhang, G.L., Sun, Y.J., Zhao, Y.G., Li, D.C., 2013. Mapping the Three-Dimensional Distribution of Soil Organic Matter across a Subtropical Hilly Landscape. Soil Sci. Soc. Am. J. 77(4), 1241-1253.
Louis, B.P., Saby, N.P.A., Orton, T.G., Lacarce, E., Boulonne, L., Jolivet, C., Ratie, C., Arrouays, D., 2014. Statistical sampling design impact on predictive quality of harmonization functions between soil monitoring networks. Geoderma 213, 133-143.
Maimon, O., Rokach, L., 2005. Data mining and knowledge discovery handbook, 2. Springer.
Maleki, M.R., van Holm, L., Ramon, H., Merckx, R., De Baerdemaeker, J., Mouazen, A.M., 2006. Phosphorus sensing for fresh soils using visible and near infrared spectroscopy. Biosystems Engineering 95(3), 425-436.
Malley, D.F., Martin, P., McClintock, L., Yesmin, L., Eilers, R., Haluschak, P., 2000. Feasibility of analysing archived Canadian prairie agricultural soils by near infrared reflectance spectroscopy, Near infrared spectroscopy: Proceeding of the 9th International Conference, Norwich, UK, NIR Publications, pp. 579-585.
Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160(3-4), 614-626.
Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154(1–2), 138-152.
Markert, B., 2007. Quality assurance of plant sampling and storage, Qualilty Assurance in Environmental Monitoring. Wiley-VCH Verlag GmbH, pp. 215-254.
Matthiesen, H., 2004. In situ measurement of soil pH. Journal of Archaeological Science 31(10), 1373-1381.
McBratney, A.B., Bishop, T.F.A., Teliatnikov, I.S., 2000a. Two soil profile reconstruction techniques. Geoderma 97(3–4), 209-221.
McBratney, A.B., Degruijter, J.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.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000b. An overview of pedometric techniques for use in soil survey. Geoderma 97(3-4), 293-327.
McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117(1-2), 3-52.
McCarty, G.W., Reeves, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Sci. Soc. Am. J. 66(2), 640-646.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21(2), 239-245.
Meersmans, J., van Wesemael, B., De Ridder, F., Van Molle, M., 2009. Modelling the three-dimensional spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 152(1–2), 43-52.
Michot, D., Walter, C., Adam, I., Guero, Y., 2013. Digital assessment of soil-salinity dynamics after a major flood in the Niger River valley. Geoderma 207, 193-204.
Miller, B.A., Schaetzl, R.J., 2015. Digital Classification of Hillslope Position. Soil Sci. Soc. Am. J. 79(1), 132-145.
Miller, C.E., 2001. Chemical principles of near-infrared technology. Near-infrared technology in the agricultural and food industries 2.
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32(9), 1378-1388.
Minasny, B., McBratney, A.B., 2007. Chapter 12 Latin Hypercube Sampling as a Tool for Digital Soil Mapping. In: A.B.M. P. Lagacherie, M. Voltz (Eds.), Developments in Soil Science. Elsevier, pp. 153-606.
Minasny, B., McBratney, A.B., 2010. Conditioned Latin Hypercube Sampling for Calibrating Soil Sensor Data to Soil Properties. Proximal Soil Sensing. Springer, Dordrecht.
Minasny, B., McBratney, A.B., Mendonça-Santos, M.L., Odeh, I.O.A., Guyon, B., 2006. Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. Soil Res. 44(3), 233-244.
Minasny, B., Stockmann, U., Hartemink, A.E., McBratney, A.B., 2016. Measuring and Modelling Soil Depth Functions. In: E.A. Hartemink, B. Minasny (Eds.), Digital Soil Morphometrics. Springer International Publishing, Cham, pp. 225-240.
Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., Van Meirvenne, M., 2009. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. Soil Sci. Soc. Am. J. 73(2), 614-621.
Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21(15), 3301-3307.
Morgan, C.L.S., Waiser, T.H., Brown, D.J., Hallmark, C.T., 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. Geoderma 151(3–4), 249-256.
Mueller, T.G., Mijatovic, B., Sears, B.G., Pusuluri, N., Stombaugh, T.S., 2004a. Soil electrical conductivity map quality. Soil Science 169(12), 841-851.
Mueller, T.G., Pusuluri, N.B., Mathias, K.K., Cornelius, P.L., Barnhisel, R.I., 2004b. Site-specific soil fertility management: A model for map quality. Soil Sci. Soc. Am. J. 68(6), 2031-2041.
Mueller, T.G., Pusuluri, N.B., Mathias, K.K., Cornelius, P.L., Barnhisel, R.I., Shearer, S.A., 2004c. Map quality for ordinary kriging and inverse distance weighted interpolation. Soil Sci. Soc. Am. J. 68(6), 2042-2047.
Mulder, V.L., de Bruin, S., Schaepman, M.E., 2013. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. International Journal of Applied Earth Observation and Geoinformation 21, 301-310.
Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263, 16-34.

Odgers, N.P., Libohova, Z., Thompson, J.A., 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma 189–190(0), 153-163.
Odgers, N.P., McBratney, A.B., Minasny, B., 2011a. Bottom-up digital soil mapping. I. Soil layer classes. Geoderma 163(1-2), 38-44.
Odgers, N.P., McBratney, A.B., Minasny, B., 2011b. Bottom-up digital soil mapping. II. Soil series classes. Geoderma 163(1-2), 30-37.
Odgers, N.P., McBratney, A.B., Minasny, B., 2015. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma 237–238(0), 190-198.
Pask, J., Turner, M., 1955. Clays and Clay Technology. Soil Science 80(1), 86.
Piikki, K., Soderstrom, M., Stenberg, B., 2013. Sensor data fusion for topsoil clay mapping. Geoderma 199, 106-116.
Piikki, K., Wetterlind, J., Soderstrom, M., Stenberg, B., 2015. Three-dimensional digital soil mapping of agricultural fields by integration of multiple proximal sensor data obtained from different sensing methods. Precision Agric 16(1), 29-45.
Poggio, L., Gimona, A., 2014. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation - An example from Scotland. Geoderma 232, 284-299.
Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 209, 1-14.
Ponce-Hernandez, R., Marriott, F.H.C., Beckett, P.H.T., 1986. An improved method for reconstructing a soil profile from analyses of a small number of samples. Journal of Soil Science 37(3), 455-467.
Qin, C.Z., Zhu, A.X., Qiu, W.L., Lu, Y.J., Li, B.L., Pei, T., 2012. Mapping soil organic matter in small low-relief catchments using fuzzy slope position information. Geoderma 171, 64-74.
Rad, M.R.P., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 232, 97-106.
Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Dematte, J.A.M., Scholten, T., 2014. Sampling optimal calibration sets in soil infrared spectroscopy. Geoderma 226, 140-150.
Reeves Iii, J., McCarty, G., Mimmo, T., 2002. The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soils. Environmental Pollution 116, Supplement 1, S277-S284.
Rizzo, R., Dematte, J.A.M., Lepsch, I.F., Gallo, B.C., Fongaro, C.T., 2016. Digital soil mapping at local scale using a multi-depth Vis-NIR spectral library and terrain attributes. Geoderma 274, 18-27.
Rossel, R.A.V., Lark, R.M., 2009. Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. European Journal of Soil Science 60(3), 453-464.
Rossel, R.A.V., Rizzo, R., Dematte, J.A.M., Behrens, T., 2010. Spatial Modeling of a Soil Fertility Index using Visible-Near-Infrared Spectra and Terrain Attributes. Soil Sci. Soc. Am. J. 74(4), 1293-1300.
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., Roudier, M.P., 2015. Package ‘clhs’.
Royle, J.A., Nychka, D., 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Comput. Geosci. 24(5), 479-488.

Russell, J.S., Moore, A.W., 1968. Comparison of different depths weightings in the numerical analysis of anisotropic soil profile data, In: Transactions of the 9th international congress of soil science, London, pp. 205-213.
Samyn, K., Cerdan, O., Grandjean, G., Cochery, R., Bernardie, S., Bitri, A., 2012. Assessment of vulnerability to erosion: Digital mapping of a loess cover thickness and stiffness using spectral analysis of seismic surface-waves. Geoderma 173, 162-172.
SAS, 2008. SAS OnlineDoc®, Version 8, Cary, NC: SAS Institute Inc., 1999.
Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry 36(8), 1627-1639.
Schmidt, K., Behrens, T., Daumann, J., Ramirez-Lopez, L., Werban, U., Dietrich, P., Scholten, T., 2014. A comparison of calibration sampling schemes at the field scale. Geoderma 232, 243-256.
Schulte, E.E., Kaufmann, C., Peter, J.B., 1991. The influence of sample size and heating time on soil weight loss-on-ignition. Communications in Soil Science and Plant Analysis 22(1-2), 159-168.
Scudiero, E., Deiana, R., Teatini, P., Cassiani, G., Morari, F., 2011. Constrained optimization of spatial sampling in salt contaminated coastal farmland using EMI and continuous simulated annealing. Spatial Statistics 2011: Mapping Global Change 7, 234-239.
Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography 27(2), 171-197.
Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66(3), 988-998.
Shonk, J., Gaultney, L., Schulze, D., Van Scoyoc, G., 1991. Spectroscopic sensing of soil organic matter content. Transactions of the ASAE 34(5), 1978-1984.
Silva, S.H.G., Owens, P.R., Silva, B.M., de Oliveira, G.C., de Menezes, M.D., Pinto, L.C., Curi, N., 2015. Evaluation of Conditioned Latin Hypercube Sampling as a Support for Soil Mapping and Spatial Variability of Soil Properties. Soil Sci. Soc. Am. J. 79(2), 603-611.
Stehman, S.V., 1999. Basic probability sampling designs for thematic map accuracy assessment. International Journal of Remote Sensing 20(12), 2423-2441.
Stehman, S.V., Czaplewski, R.L., 1998. Design and analysis for thematic map accuracy assessment: Fundamental principles. Remote Sensing of Environment 64(3), 331-344.
Stenberg, B., Jonsson, A., Börjesson, T., 2002. Near infrared technology for soil analysis with implications for precision agriculture, Near Infrared Spectroscopy: Proceedings of the 10th International Conference, Kyongju S. Korea. NIR Publications, Chichester, UK, pp. 279-284.
Stenberg, B., Viscarra Rossel, R.A., 2010. Diffuse reflectance spectroscopy for high-resolution soil sensing, Proximal Soil Sensing. Springer, pp. 29-47.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter Five - Visible and Near Infrared Spectroscopy in Soil Science. In: L.S. Donald (Ed.), Advances in Agronomy. Academic Press, pp. 163-215.
Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods 14(4), 323.
Sun, W., Minasny, B., McBratney, A., 2012a. Analysis and prediction of soil properties using local regression-kriging. Geoderma 171–172, 16-23.
Sun, X.L., Wu, S.C., Wang, H.L., Zhao, Y.G., Zhao, Y.C., Zhang, G.L., Man, Y.B., Wong, M.H.,

2012b. Uncertainty Analysis for the Evaluation of Agricultural Soil Quality Based on Digital Soil Maps. Soil Sci. Soc. Am. J. 76(4), 1379-1389.
Taghizadeh-Mehrjardi, R., 2016. Digital mapping of cation exchange capacity using genetic programming and soil depth functions in Baneh region, Iran. Arch. Agron. Soil Sci. 62(1), 109-126.
Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., Malone, B.P., 2014. Digital mapping of soil salinity in Ardakan region, central Iran. Geoderma 213(0), 15-28.
Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma 266, 98-110.
Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253, 67-77.
Terra, F.S., Demattе, J.A.M., Rossel, R.A.V., 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis-NIR and mid-IR reflectance data. Geoderma 255, 81-93.
Theocharopoulos, S.P., Wagner, G., Sprengart, J., Mohr, M.E., Desaules, A., Muntau, H., Christou, M., Quevauviller, P., 2001. European soil sampling guidelines for soil pollution studies. Science of The Total Environment 264(1–2), 51-62.
Thomas, M., Clifford, D., Bartley, R., Philip, S., Brough, D., Gregory, L., Willis, R., Glover, M., 2015. Putting regional digital soil mapping into practice in Tropical Northern Australia. Geoderma 241, 145-157.
Thomas, M., Odgers, N.P., Ringrose-Voase, A., Grealish, G., Glover, M., Dowling, T., 2012. Soil survey design for management-scale digital soil mapping in a mountainous southern Philippine catchment, Digital Soil Assessments and Beyond. CRC Press, pp. 233-238.
Triantafilis, J., Earl, N.Y., Gibbs, I.D., 2012. Digital soil-class mapping across the Edgeroi district using numerical clustering and gamma-ray spectrometry data, Digital Soil Assessments and Beyond. CRC Press, pp. 187-191.
van Groenigen, J.W., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87(3–4), 239-259.
van Groenigen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. J. Environ. Qual. 27(5), 1078-1086.
van Zijl, G.M., Bouwer, D., van Tol, J.J., le Roux, P.A.L., 2014. Functional digital soil mapping: A case study from Namarroi, Mozambique. Geoderma 219, 155-161.
Vašát, R., Heuvelink, G.B.M., Borůvka, L., 2010. Sampling design optimization for multivariate soil mapping. Geoderma 155(3–4), 147-153.
Vašát, R., Lubo, B.v., Ond_ej, J.í., 2012. Number of sampling points influences the parameters of soil properties spatial distribution and kriged maps, Digital Soil Assessments and Beyond. CRC Press, pp. 251-256.
Vasques, G.M., Demattê, J.A.M., Viscarra Rossel, R.A., Ramírez López, L., Terra, F.S., Rizzo, R., De Souza Filho, C.R., 2015. Integrating geospatial and multi-depth laboratory spectral data for mapping soil classes in a geologically complex area in southeastern Brazil. European Journal of Soil Science 66(4), 767-779.
Vasques, G.M., Grunwald, S., Comerford, N.B., Sickman, J.O., 2010. Regional modelling of soil carbon at multiple depths within a subtropical watershed. Geoderma 156(3-4), 326-336.
Vaysse, K., Lagacherie, P., 2015. Evaluating Digital Soil Mapping approaches for mapping

GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Regional 4(0), 20-30.
Vela, N., Hernandez, J., Vicente, A.P.M., Ortiz, R., 2005. Spatial variability of salinity in cultivated soils of semiarid Murcia, SE Spain. In: A.F. Cano, R.O. Silla, A.R. Mermut (Eds.), Sustainable Use and Management of Soils - Arid and Semiarid Regions. Advances in Geoecology. Catena Verlag, Reiskirchen, pp. 545-558.
Vendrame, P.R.S., Marchão, R.L., Brunet, D., Becquer, T., 2012. The potential of NIR spectroscopy to predict soil texture and mineralogy in Cerrado Latosols. EJSS European Journal of Soil Science 63(5), 743-753.
Venerables, W., Ripley, B., 2002. Modern Applied Statistics with S. new york: Springer.
Veronesi, F., 2012. 3D Advance mapping of soil properties, CRANFIELD UNIVERSITY.
Veronesi, F., Corstanje, R., Mayr, T., 2012. Mapping soil compaction in 3D with depth functions. Soil and Tillage Research 124(0), 111-118.
Viscarra Rossel, R.A., 2008. ParLeS: Software for chemometric analysis of spectroscopic data. Chemometrics and Intelligent Laboratory Systems 90(1), 72-83.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158(1-2), 46-54.
Viscarra Rossel, R.A., Cattle, S.R., Ortega, A., Fouad, Y., 2009. In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy. Geoderma 150(3-4), 253-266.
Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Res. 53(8), 845-864.
Viscarra Rossel, R.A., McBratney, A.B., 1998. Soil chemical analytical accuracy and costs: implications from precision agriculture. Australian Journal of Experimental Agriculture 38(7), 765-775.
Viscarra Rossel, R.A., McGlynn, R.N., McBratney, A.B., 2006a. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. Geoderma 137(1–2), 70-82.
Viscarra Rossel, R.A., Minasny, B., Roudier, P., McBratney, A.B., 2006b. Colour space models for soil science. Geoderma 133(3–4), 320-337.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006c. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131(1–2), 59-75.
Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T., 2007. In Situ Characterization of Soil Clay Content with Visible Near-Infrared Diffuse Reflectance Spectroscopy. Soil Sci. Soc. Am. J. 71(2), 389-396.
Walvoort, D.J.J., Brus, D.J., de Gruijter, J.J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Comput. Geosci. 36(10), 1261-1267.
Wang, D.C., Zhang, G.L., Rossiter, D.G., Zhang, J.H., 2016. The Prediction of Soil Texture from Visible-Near-Infrared Spectra under Varying Moisture Conditions. Soil Sci. Soc. Am. J. 80(2), 420-427.
Wang, Y., Huang, T., Liu, J., Lin, Z., Li, S., Wang, R., Ge, Y., 2015. Soil pH value, organic matter and macronutrients contents prediction using optical diffuse reflectance spectroscopy. Computers and Electronics in Agriculture 111, 69-77.

Webster, R., 1978. Mathematical treatment of soil information, Transactions of the 11th International Congress of Soil Science, pp. 161-190.
Webster, R., Lark, M., 2013. Field sampling for environmental science and management, 14. Springer US.
Wei, T., 2013. corrplot: visualization of a correlation matrix. R package version 0.73.
Wheeler, I., McBratney, A.B., Minasny, B., Gruijter, J.J.d., 2012. Digital Soil Mapping to inform design-based sampling strategies for estimating total organic carbon stocks at the farm scale, Digital Soil Assessments and Beyond. CRC Press, pp. 257-262.
White, K., Walden, J., Drake, N., Eckardt, F., Settlell, J., 1997. Mapping the iron oxide content of dune sands, Namib Sand Sea, Namibia, using Landsat Thematic Mapper data. Remote Sensing of Environment 62(1), 30-39.
Wiese, L., Ros, I., Rozanov, A., Boshoff, A., de Clercq, W., Seifert, T., 2016. An approach to soil carbon accounting and mapping using vertical distribution functions for known soil types. Geoderma 263, 264-273.
Williams, B.K., 1996. Assessment of accuracy in the mapping of vertebrate biodiversity. Journal of Environmental Management 47(3), 269-282.
Worsham, L., Markewitz, D., Nibbelink, N.P., West, L.T., 2012. A Comparison of Three Field Sampling Methods to Estimate Soil Carbon Content. For. Sci. 58(5), 513-522.
Wu, C.-Y., Jacobson, A.R., Laba, M., Kim, B., Baveye, P.C., 2010. Surrogate correlations and near-infrared diffuse reflectance sensing of trace metal content in soils. Water, Air, & Soil Pollution 209(1-4), 377-390.
Zhang, J., Zhang, C., 2011. Sampling and sampling strategies for environmental analysis. International Journal of Environmental Analytical Chemistry 92(4), 466-478.
Zhu, Q., Lin, H., 2010. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. Pedosphere 20(5), 594-606.
Ziadi, N., Tran, T., 2007. Mehlich 3-extractable elements. Soil sampling and methods of analysis. Lewis, Boca Raton, FL, 81-88.

PREFACE TO APPENDIX A

Soil profile depth function is essential for quantifying vertical variability of soil properties and assisting in the three-dimensional regression kriging. The most accurate and flexible profile depth function is the equal-area quadratic spline function (EAQSF), which has been widely used in 3D-DSM in literatures and in previous chapters of this thesis. However, one limitation of EAQSF is that it is a simple mathematical and graphical fitting of the vertical distribution of soil properties without considering any physical conditions of soil. Therefore, more specific depth function with interpretation of soil formation and morphology should be searched. Exponential decay function is an example which was highly developed for modelling vertical distribution of soil organic matter. However, its feasibility for other soil properties was doubted. In addition, other mathematical functions, such as polynomial function and power function, were also occasionally used for case studies, but not widely proved and applied. Therefore, based on understanding the physical conditions of the agricultural field, we proposed a new model to predict soil pH at depths, and tested its generality for a global dataset with 432 soil profiles in the appendix A. However, this newly proposed model was not used in 3D mapping of soil pH, while it could be further assessed in future research. A manuscript was formatted with the name 'A sigmoid depth function to describe variations in soil pH in agricultural fields' was revised and resubmitted to *Geoderma* (impact factor 2.85) on Jun. 17th 2016. The detailed information for authors is shown below:

Order of authors: Yakun Zhang[a], Asim Biswas[a,*], & Viacheslav I. Adamchuk[b].

Contribution of authors: The author of the thesis was responsible for the laboratory measurement, data analysis and interpretation, and manuscript preparation. Prof. Biswas proposed the new model and assist in all the work. Prof. Adamchuk gave much suggestions on the development and analysis of the model.

APPENDIX A- A sigmoid depth function to describe variations in soil pH in agricultural fields

Abstract:

Soil pH controls the availability of the majority of plant nutrients, if not all, and determines the growth environment for plant roots. Profile depth functions have been used to represent the vertical distribution of soil attributes and to predict them at continuous depths. This paper proposes a new model to predict pH for a whole soil profile. Soil properties including pH are often similar within the plough layer from mixing during tillage and other agricultural operations. Similarly, soil pH below the root zone tends to be very uniform due to least disturbance, leaving a transition zone from the bottom of the tillage layer to the bottom of the root zone with variable pH contributed from variable root density and activity. Keeping this physical description of agricultural field soil profile in mind, a closed form equation (model) was developed similar to a sigmoid curve. The model has 4 parameters including 1) soil pH at the top of a soil profile, 2) soil pH at the bottom of a soil profile, 3) hillslope parameter representing steepness of the curve that is determined by the length of the root zone, and 4) inflection point representing almost the midpoint of the transition zone or root zone. A total of 32 soil cores down to about 1.1 m depths were collected from an agricultural field of Macdonald farm, McGill University. The sub-samples were taken at every 10 cm and analyzed for pH in the laboratory in soil: water suspension. The lab measured pH was used to test the fitting performance of the sigmoid model. Additionally, a global dataset with 432 profiles with various soil classes, drainage types, land use, and altitude was also used to test the generality of the new model. The performance of this model was compared with the results of the commonly used 3rd order polynomial regression function and the equal-area quadratic spline function. Good performance of the sigmoid model with explicit physical explanation showed promise in predicting soil pH at depths. The spline function had the highest accuracy but lacked a general trend in its shape and parameters. The polynomial function had good accuracy and displayed a non-monotonic trend, which can also be used as a substitute for some profiles with complex variability.

Keywords: depth function, sigmoid model, soil pH, digital soil mapping, 3D variability, agricultural soil

A.1 Introduction

Soil pH is an important soil quality index and controls the plant nutrient availability, growth environment of plant roots, soil microbial activities, and many chemical processes that take place

in soil (Aciego Pietri and Brookes, 2008; Kahlert et al., 2004). Agricultural management decisions are often constrained to surface soil pH measurements due to the convenience and ease of sample collection. However, as the plant roots can reach subsoil and even deep soil, the measurement of soil pH at depths is important for understanding the rhizosphere environment, and chemical and biological activities. Various soil forming factors (Jenny, 1941), such as parent material, organism, and climate, combined with management activities, like fertilizer and manure application, and tillage, contribute together to the variability of soil pH. Quantitative information on the spatial variability of soil pH, sometimes displayed as digital soil maps, plays an essential role in site-specific agricultural management such as lime requirements, and soil quality assessment. Additionally, 3D digital soil mapping combining horizontal maps and profile depth functions becomes increasingly popular and important for understanding three-dimensional spatial variability and its relationship to other soil properties (Liu et al., 2013).

Profile depth functions are based on the premise that soil properties vary continuously with depths in a profile (Russell and Moore, 1968). The variability has been modelled by various depth functions, ranging from a freehand curve created by Jenny (1941), to more sophisticated models, such as exponential decay functions (EDFs) (Minasny et al., 2006), polynomial functions (Veronesi et al., 2012), power functions (Liu et al., 2016), and equal-area quadratic spline functions (EAQSFs) (Bishop et al., 1999). The EAQSFs, fitted by a set of local quadratic polynomials for each horizon, describe a smooth curve through horizon mid-points (McBratney et al., 2000a). Spline functions were reported to have the highest accuracy due to higher flexibility and feasibility (Webster, 1978). The EAQSF has been widely and successfully used to model the vertical distribution of various soil properties, including soil organic carbon, available water capacity, soil texture, bulk density, soil salinity, and soil pH (Adhikari et al., 2014; Bishop et al., 2015; Lacoste et al., 2014; Malone et al., 2009; Odgers et al., 2012; Taghizadeh-Mehrjardi et al., 2014). However, without an explicit mathematical formula and a consistent set of parameters, EAQSF is simply a numerical and graphical fitting of horizon data, which changes its shape from profile to profile. Such function individually fits the profile data well but lacks a general trend and the physical explanation of soil-landscape relationship (Liu et al., 2016). Therefore, more effective depth functions with definite mathematical formulas, and clear and general tendency should be searched for specific soil properties.

The EDFs have been used to model the vertical distribution of soil organic carbon content (SOC)

basing the fact that higher SOC is present in the topsoil and gradually decreases in the profile (Minasny et al., 2006). Later the EDFs have been modified by involving the integral form (Mishra et al., 2009), segmenting the functions with a constant presenting plough layers (Kempen et al., 2011; Meersmans et al., 2009), and creating a normalized form (Wiese et al., 2016) to take into account practical issues and represent site-specific profiles. However, the monotonic and steady decreasing trends of the EDFs limit their application for other soil properties. In recent years, more and more mathematical models are proposed to delineate the vertical distribution of various soil properties, including a 6^{th} order polynomial regression functions to represent soil compaction (Veronesi et al., 2012), a linear function with Tikhonov regularization (TR) to describe soil EC (Li et al., 2013), and another power function to describe SOC (Liu et al., 2016). Moreover, Minasny et al. (2016) reviewed several common types of parametric and nonparametric depth functions, including uniform, gradational, exponential, wetting front, abrupt, peak, and MiniMax; some of which only have graphical fitting and lack mathematical formulas. Even though these depth functions fit well, the generality of these functions still need further exploration, and the physical explanation of the parameters needs improvement to represent the effect pf pedological process and management activities.

Every soil property has its unique vertical distribution which could be modeled by specific depth function (Jenny, 1941). However, soil pH has not been widely recognized for its vertical variability and modelled by explicit equations. Yet few papers used data or graphs to qualitatively show the vertical trend of pH values. For example, Chi et al. (2010) reported an increasing soil pH with depth in reclaimed rice land and soybean land. The EAQSFs have also been used to model the vertical distribution of soil pH for digital soil mapping (Adhikari et al., 2012; Bishop et al., 1999; Odgers et al., 2015). However, the EAQSF fits soil profile individually and lacks generality. Moreover, considering the physical condition of agricultural fields, three types of variability in soil pH may persist with depth: 1) a relatively uniform condition in the plough layer due to the mixing effect of tillage and other agricultural operations, 2) a relatively uniform condition in the bottom layer due to non-disturbance and possible consistent groundwater effect, and 3) a transition layer in between. Soil pH should be fitted with a more general and appropriate function that can better describe the variability with depth.

The main objective of this paper is to develop a new closed form sigmoid model and test its ability in predicting soil pH in agricultural fields. More specifically, the paper aims: 1) to develop and test

a sigmoid model in predicting soil pH in a small agricultural field (using a local dataset); 2) to fit the model with a global soil pH dataset to test the universality of the sigmoid model in predicting soil pH; and 3) to compare the predictive capability of the new model with the commonly used 3rd order polynomial regression function and EAQSF.

A.2 Materials and methods

A.2.1 Study area

A field experiment was conducted in Field 26 (11 ha) of Macdonald Farm, McGill University, Quebec, Canada (45.4° N and 73.9° W) (Fig. A.1). The landscape of the farm locates on two rolling plateaus formed by thousands of years' carving of Ottawa River, resulting in various soil types and providing a good test bed for model validation. Soil types of Field 26 are highly variable and range from the deep organic deposit (peat) over the shallow organic deposit to mineral soils with dominant textures including sand, light sandy loam, ill-drained sandy loam, loam, silt loam, and clay loam. Soils in Field 26 are classified into multiple soil series including Muck, ST-Zotique, Soulanges, ST-Damase, Uplands, Chicot, Farmington, Chateauguay, and Macdonald following the Canadian Soil Classification System. The elevation of Field 26 ranges from 6.88 to 9.22 meters above sea level and the long-term (30 years) average annual air temperature is 6.2°C and average annual precipitation is 979 mm. Field 26 was under corn-soybean rotation and the crop previous to sample collection was soybean.

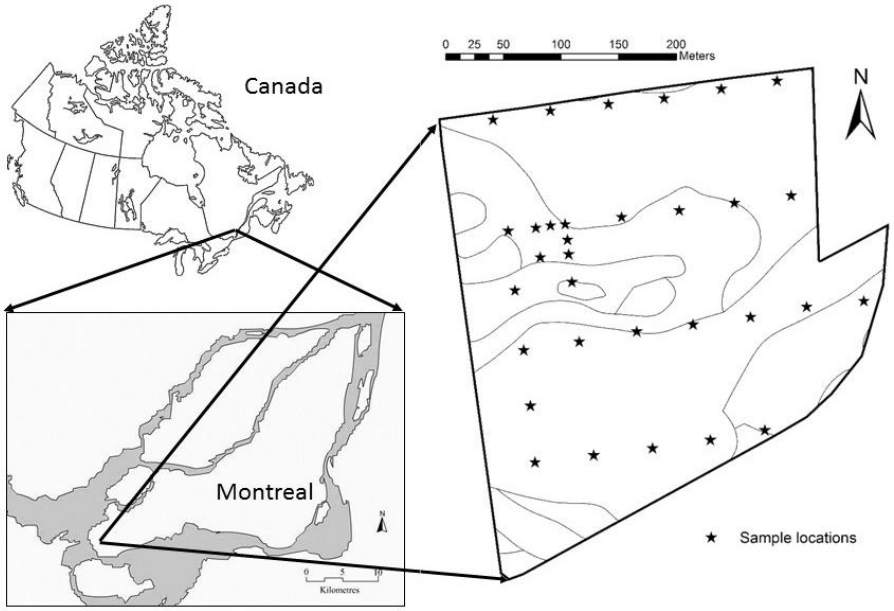

Fig. A.1. Study area at the Macdonald farm of McGill University, Montreal, Canada showing the sample locations in Field 26.

A.2.2 Sample collection and analysis

A total of 32 georeferenced soil cores (Fig. A.1) down to about 1.1 m depth were collected using a truck-mounted hydraulic soil profiler (Veris® P4000 soil profiler, Veris technologies Inc., Salina, KS, USA) following a modified nested grid sampling design to obtain a good spatial coverage in November 2014. The soil cores were subsampled at every 10 cm layer. Two soil profiles were dug only to 30 cm restricted by rocks occurring at a shallow depth. A total of 284 samples were sealed in Ziploc bags and transported to the laboratory for analysis.

Air-dried and ground (particle size < 2 mm) samples were used for soil pH determination in a soil-water solution of 1:2 soil to water ratio (1:4 for organic soil). Since the samples were taken at 10-cm depth intervals, the measured pH values represented the average values of 10 cm soil horizons and marked as the pH value at mid-point of each soil horizon.

A.2.3 Sigmoid model

A new sigmoid model was adopted in this research as follows:

$$f(x) = \mathrm{s} + \frac{\mathrm{d}-\mathrm{s}}{1+(\frac{x}{\alpha})^{-k}} \quad \text{(A.1)}$$

where $f(x)$ was the soil pH, and x was the soil depth. s and d represented the soil pH at the top and bottom of soil profiles, respectively, α was the inflection point which represented almost the midpoint of the transition zone and k represented the steepness of the curve which returned the largest absolute value of the slope of the curve. A possible approach to estimating s and d is to impose the two parameters in the sigmoid function by measured values of soil pH at the top and bottom of soil profiles.

A.2.4 Depth functions

The sigmoid model was compared with the commonly used 3rd order polynomial regression function and the EAQSF.

For 3rd order polynomial regression function used in this study was:

$$f(x) = a + b \times x + c \times x^2 + d \times x^3 \quad \text{(A.2)}$$

where a, b, c, and d were four parameters of the polynomial function. The 3rd order polynomial function was chosen in this study because it had four parameters which made it comparable to four-parameter sigmoid function. The sigmoid and 3rd order polynomial functions were fitted by minimizing RMSE with 'fminsearchbnd' function in MATLAB (The MathWorks Inc., Release: R2015b).

A detailed description of the application of the EAQSF can be found in Bishop et al. (1999). Briefly,

it is assumed that the bulk soil attribute represented the mean of the soil horizon and had two key characteristics: 1) it consisted of a series of local quadratic functions with 'knots' at the boundaries of the horizons; 2) for each horizon, the area to the left of the fitting curve was equal to the area to the right of the fitting curve (Ponce-Hernandez et al., 1986). Basically, the spline was achieved by minimizing the function:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{f_i})^2 + \lambda \int_{x_0}^{x_n} f'(x)^2 dx \quad \text{(A.3)}$$

The first part of the equation determined the goodness-of-fit, and the second part determined the roughness. The λ controlled the trade-off between the goodness-of-fit and the roughness and the λ values of 10, 1, 0.1, 0.01, and 0.001 were tested in this case to examine the best fit. The EAQSFs were fitted using code written in MATLAB R2015b.

A.2.5 Global dataset

A total of 432 soil profiles from agricultural fields across the world were selected from 'The International Soil Reference and Information Centre-World Inventory of Soil Emission Potentials (ISRIC-WISE)' international soil profile dataset (Batjes, 2000) to test the universality of the sigmoid profile depth function. All georeferenced profiles were classified according to the 1974 Legend of the FAO-UNESCO Soil Map (FAO-UNESCO, 1974) of the World, as well as the 1988 Revised Legend of FAO-UNESCO (FAO, 1990). Soil pH values measured in 1:2.5 soil to water solution were selected for this study. Depth intervals were not uniform for all the profiles. The selected dataset came from various regions, crop types, climate, parent materials, thus allowing us to comprehensively test the sigmoid model. The fitting results of various soil class, land use, drainage, and altitude were compared to interpret the effects of various pedological and environmental conditions on the soil pH values and the performance of the sigmoid model.

A.2.6 Accuracy and efficiency

The predictive quality of these depth functions was determined by comparing their estimated value (f_i) and the true value (y_i) obtained by laboratory analysis. The common statistical indices, including root mean squared error (RMSE) and the coefficient of determination (R^2) were used to test the prediction quality. The RMSE measured the average magnitude of errors between the predicted and measured values and indicated the accuracy of prediction (Liu et al., 2013). The R^2 indicated the effectiveness when using one variable to predict another variable (Taghizadeh-Mehrjardi, 2016). Basically, a function with larger R^2 (close to 1) and smaller RMSE (close to 0) was regarded as the good fitting scheme.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2} \quad \text{(A.4)}$$

$$\mathrm{R}^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{(A.5)}$$

where n was the number of samples and SS_{res} and SS_{tot} were the sum of squared error of residuals and the total, respectively.

A.3 Results and discussions

A.3.1 Profile description

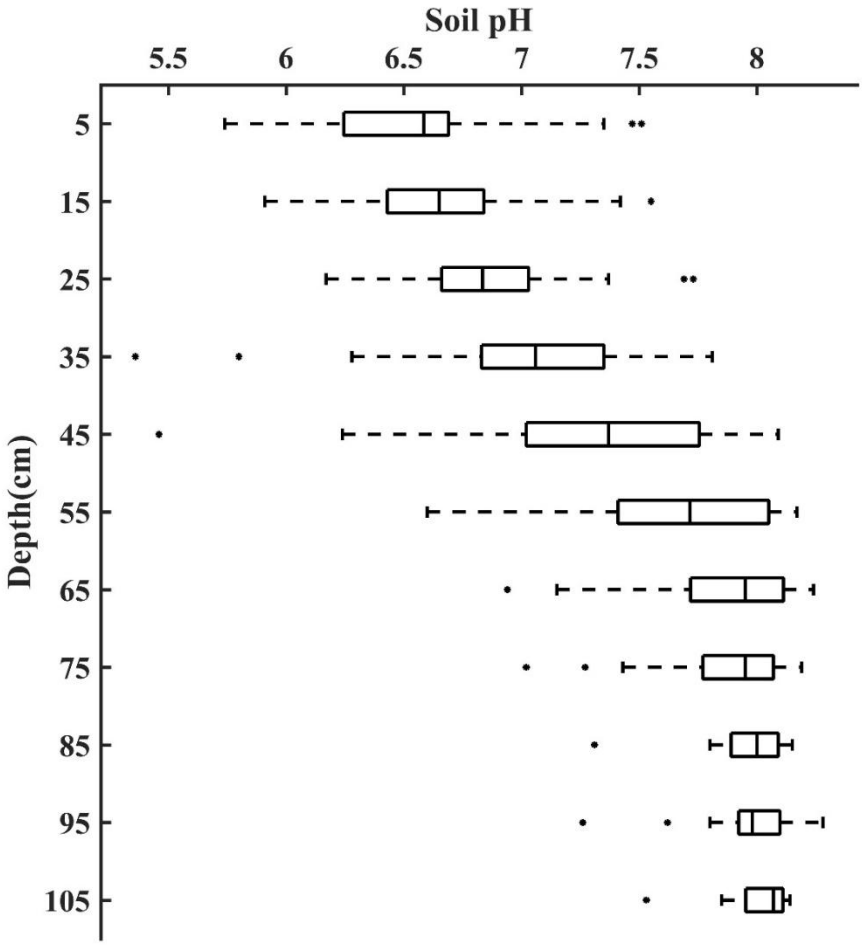

Fig. A.2. Box Plot of variation of soil pH with depth in 32 profiles collected from the study site.

The soil pH values ranged from 5.36 to 8.28, with 99% between the range from slightly acidic to moderately alkaline (6.1-8.4) and were regarded as a good condition for crop growth (Hazelton and Murphy, 2007). An increasing trend of the average values was observed with depth (Fig. A.2), with a relatively uniform values of soil pH in top three layers (mean values ranging from 6.56 to 6.86) and in the deepest five layers (mean values ranging from 7.85 to 8.01), and a transition zone

in subsoil layers (mean values ranging from 7.04 to 7.64). The vertically increasing tendency was also reported in soil pH maps (Odgers et al., 2015) and in profile descriptions (Chi et al., 2010). However, a decreasing trend was also reported by Adhikari et al. (2012) which was assumed to be the artificial influence of lime application in plow layers. In addition, a relatively large range of soil pH values was observed in transition zone with higher coefficient of variations (CV) values (0.8 and 0.6), compared to smaller range and lower CV values in top soil (0.6 and 0.5), and deep soil (0.4, 0.3, and 0.2) respectively (Fig. A.2). The negatively skewed distributions (negative values of skewness) at deeper layers indicated a number of higher values of pH at the bottom of the soil profiles.

The vertical distribution can be explained by the physical condition of this agricultural fields. Lower pH in surface layers may be due to the decay of high amount of soil organic matter (SOM) and release of weak organic acids. The SOM content of top three layers in most sampling points was between 20 and 30%. The water-saturated condition of Gleysolic soil (Canadian System of Soil Classification) – a dominant soil type of this agricultural field reduced the rates of decomposition resulting in the SOM build-up in the surface layers. Another reason for lower soil pH may be due to the removal of bases by high-yielding crops (Fernández and Hoeft, 2009) and high rainfall in the area (annual average 979 mm). In addition, mixing effect of tillage created a uniform soil pH in tillage layers (Adhikari et al., 2012). The homogeneity of other soil properties in plow layers was also reported by many others pointing towards the effect of tillage in agricultural fields (Kempen et al., 2011; Liu et al., 2016; Meersmans et al., 2009).

Conversely, the higher soil pH in bottom soil layers was due to the lower SOM, minimal management disturbance, and the influence of fluctuating ground water (study area location on Montreal Island) and the pedological process. High amount of Mg and Ca content (data not shown in this paper) in bottom layers compared to the transition zone also influenced the pH. Moreover, the presence of mottles in many profiles (identified during soil classification) indicated a fluctuating water table and the influence of groundwater on soil pH. These effects were not highly variable and yielded similar values of soil pH in deep layers. The transition zone was from the bottom of the tillage layers to the bottom of the root zone, where soil pH changed gradually as the root density and activity changed. Outliers with extremely low soil pH in the transition zone were from organic soil profiles. This localized variations may be due to the complex decomposition process of SOM and the acids formed during the process.

A.3.2 Sigmoid models in local dataset

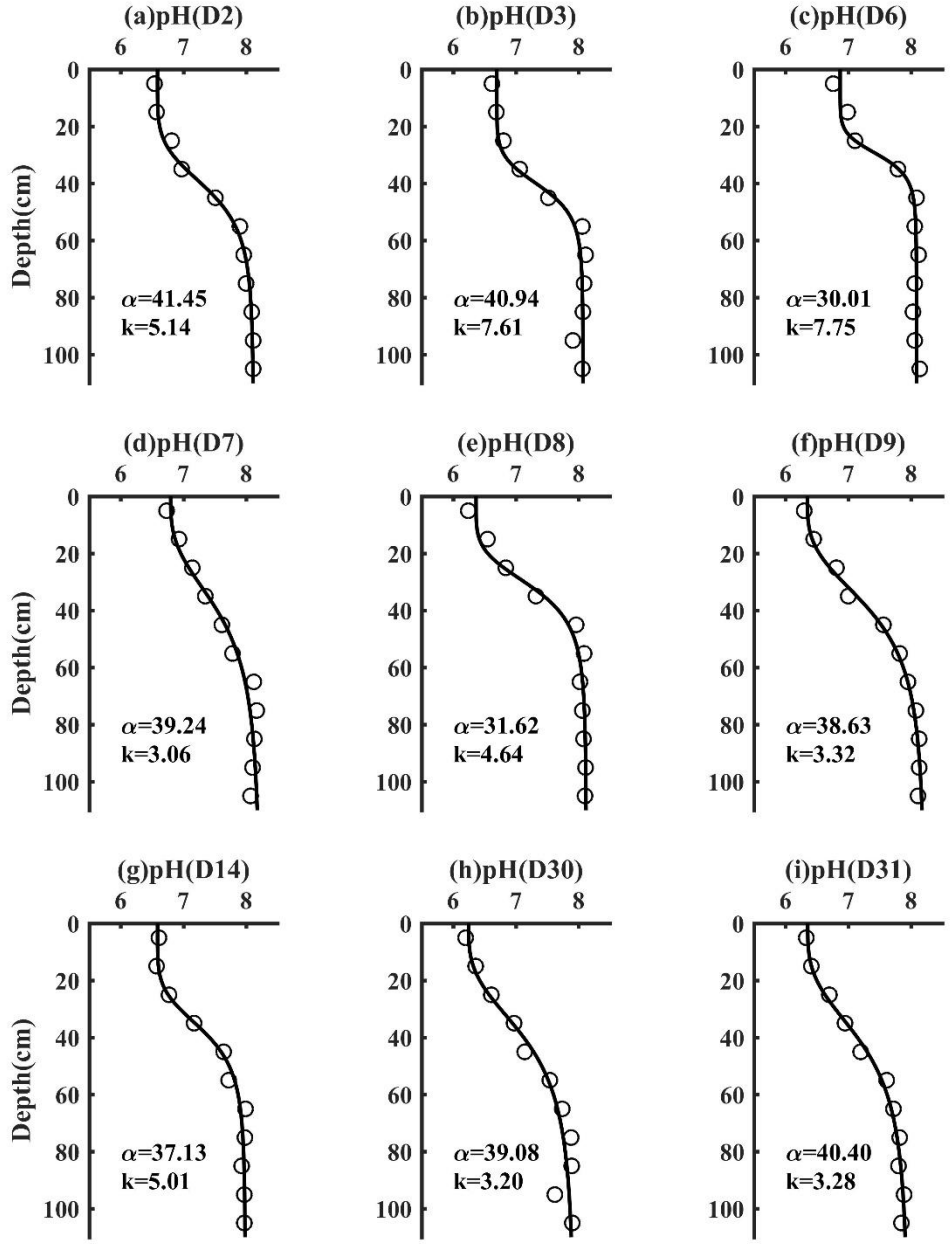

Fig. A.3. Fitted sigmoid model for nine soil profiles.

The sigmoid model was individually fitted to pH values of different depths in each soil profile and few examples were presented in Fig. A.3. From the overall distribution of soil pH (Fig. A.2) and individually fitted model values (Fig. A.3), it was clear that the vertical distribution of soil pH formed a pattern with uniform soil pH in the top and bottom layer and the transition zone in between. This pattern can be explained by the physical condition of the agricultural field as a result of a series of human and natural processes as discussed in section A.3.1. A typical soil profile with

identified horizons and corresponding sigmoid model was shown in Fig. A.4. The profiles in local dataset generally had a uniform Ap horizon of about 20-25 cm which was modelled by surface soil pH (*s*). In addition, the deep soil horizons (C horizons) below 60 cm also showed uniform pattern which was modelled by deep soil pH (*d*). A transition zone (mainly B horizons) was represented by the inflection point (α) and steepness parameters (k). Once the optimal parameters were identified, the model was used to predict unmeasured values of any depth of the profile.

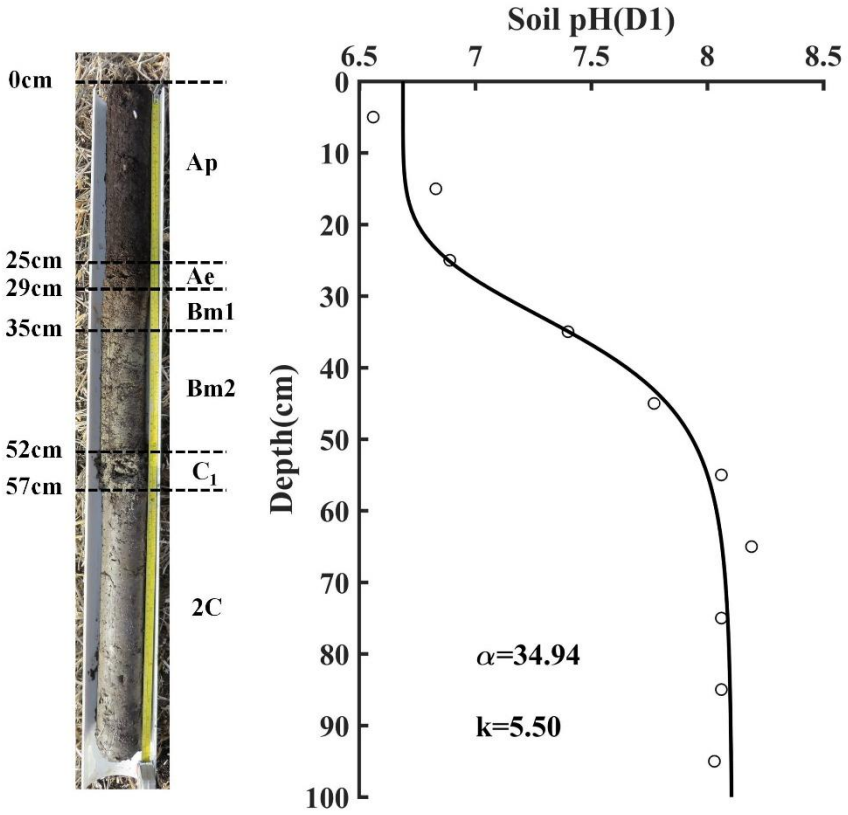

Fig. A.4. A typical soil profile of field 26 and corresponding sigmoid model. The horizons were classified following Canadian System of Soil Classification.

The measured and predicted soil pH values of the sigmoid models closely spread along 1:1 line and indicated a good fitting performance (Fig. A.5a). In addition, majority of the individually fitted profile functions showed a smaller RMSE (the mean value of RMSE=0.08 with standard deviation (sd)=0.08) and larger R^2 (the mean value of R^2=0.97 with sd=0.05) between the measured and predicted soil pH of all the soil profiles (Fig. A.5b). However, a poor performance of sigmoid model with larger RMSE values and smaller R^2 values was present in organic soil profiles (Fig. A.5b). For example, the RMSE values of 0.27 and 0.41 and R^2 values of 0.79 and 0.79 were observed in two organic soil profiles (D24 and D32, respectively). The predicted values were

exactly the same in the top layers and in the bottom layers, whereas the real conditions of these profiles were more complex with extremely low pH values in the subsoil layers owing to the various decomposition degrees and products. The sigmoid model was not able to include the complexity and changeability of organic soil profiles. Samples of relatively low pH value which were far from the fitting line in Fig. A.4a were mainly the values of organic soil profiles. The fitting results of two shallowest soil profiles (D17 and D18) with only three soil layers (up to 30 cm depth) were surprisingly good with RMSE values close to zero and R^2 values close to 1. Generally, fitting a four parameter model requires, at least, four or more data to explain the complexity, whereas shallow profiles had simpler distribution tendencies and were easily delineated by mathematical formulas. Some unreasonable values were observed during the sigmoid function fitting process mainly due to over-fitting which was solved by imposing appropriate constraints at the beginning and through iterative fitting processes with different starting values.

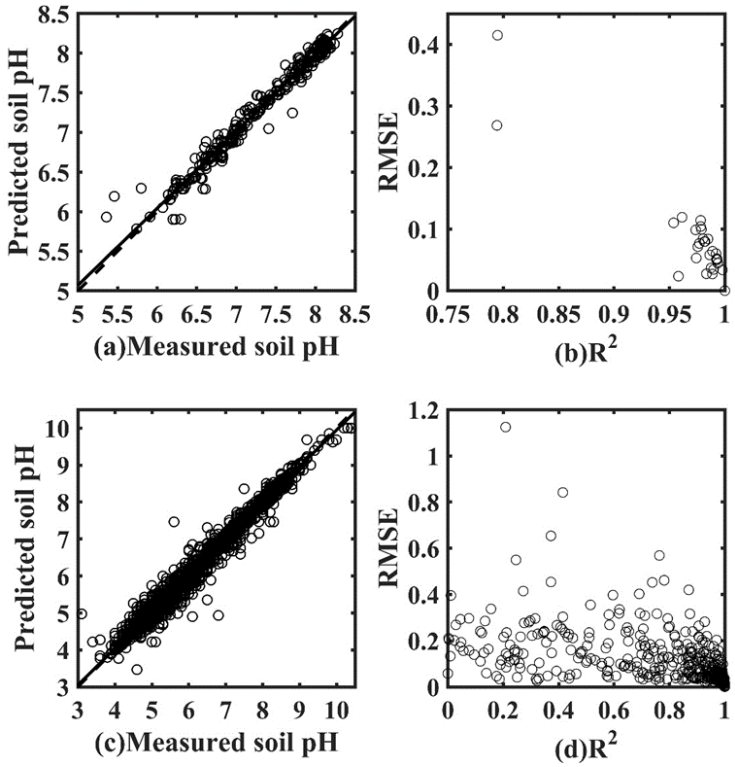

Fig. A.5. Fitting performance of sigmoid model (a) comparison of measured and predicted soil pH values of all the data points in local dataset. Dashed line is 1:1 line, and solid line is fitted by pH data. (b) Scatterplot of R2 and RMSE of every profile in local dataset. (c) Comparison of measured and predicted soil pH values of all the data points in global dataset. Dashed line is 1:1 line, and solid line is fitted by pH data. (d) Scatterplot of R2 and RMSE of every profile in global dataset.

A.3.3 Sigmoid models in global dataset

In order to test the generality of the sigmoid model, the global dataset with 432 soil profiles was individually fitted with sigmoid function and fitting results were shown in Fig. A.5. The mean value of RMSE were 0.11 with the standard deviation of 0.12, slightly larger than the local dataset (RMSE=0.08). In contrast, the mean value of R^2 were 0.76 with the standard deviation of 0.29, remarkably lower value than for local dataset (R^2= 0.97). The variability in the soil types, climate, parent materials, terrain attributes, and crop types led to more complicated and variable profile conditions and soil properties in the global dataset. Additionally, the measurement uncertainty also contributed to the variability. Therefore, the fitting results were not as good as those for the local dataset. Yet, keeping all these background conditions in mind, the RMSE, and R^2 value still indicated a moderate fitting result. Furthermore, the fitting results of the sigmoid model were further categorized into 27 soil classes, 12 kinds of land uses, 7 kinds of drainage conditions, and 9 ranges of altitude (Table A.1). A range of performances were observed for different groups of soil profiles within categorized soil types. With unequal division, the number of profiles within each group could influence on the uncertainty and performances. In spite of all these, a moderate performance of the sigmoid model fitting clearly showed promise as a new depth function to predict soil pH in agricultural fields at depths.

Table A.1 Fitting results of sigmoid model by considering soil type, land use, drainage, and altitude

KEY	Soil groups	Count	RMSE(sd)	R^2(sd)
AC	Acrisols	32	0.14(0.11)	0.73(0.29)
AL	Alisols	8	0.12(0.09)	0.54(0.39)
AN	Andosols	21	0.10(0.09)	0.81(0.29)
AR	Arenosols	4	0.11(0.14)	0.87(0.21)
AT	Anthrosols	5	0.04(0.03)	0.63(0.27)
CH	Chernozems	1	0.03(0)	0.97(0)
CL	Calcisols	8	0.09(0.06)	0.76(0.25)
CM	Cambisols	71	0.09(0.08)	0.79(0.23)
FL	Fluvisols	22	0.16(0.25)	0.63(0.37)
FR	Ferralsols	46	0.10(0.07)	0.71(0.31)
GL	Gleysols	22	0.12(0.13)	0.82(0.21)
GR	Greyzems	1	0.17(0)	0.90(0)
GY	Gypsisols	3	0.001(0.001)	0.999(0.001)
KS	Kastanozems	1	0.004(0)	0.999(0)
LP	Leptosols	2	0.08(0.04)	0.85(0.08)
LV	Luvisols	40	0.16(0.14)	0.77(0.27)
LX	Lixisols	19	0.11(0.10)	0.75(0.30)
NT	Nitisols	4	0.11(0.08)	0.76(0.19)
PD	Polzoluvisols	1	0.13(0)	0.72(0)
PH	Phaeozems	46	0.11(0.14)	0.78(0.27)

PL	Planosols	3	0.19(0.11)	0.62(0.42)
PT	Plinthosols	5	0.22(0.08)	0.50(0.24)
PZ	Podzols	3	0.16(0.06)	0.56(0.32)
RG	Regosols	8	0.08(0.04)	0.71(0.27)
SC	Solonchaks	3	0.07(0.003)	0.67(0.27)
SN	Solonetzes	8	0.17(0.13)	0.80(0.23)
VR	Vertisols	45	0.09(0.07)	0.82(0.22)
KEY	Land use	Count	RMSE(sd)	R^2(sd)
A	Crop agriculture	266	0.11(0.11)	0.75(0.27)
AA	Annual field cropping	1	0.06(0)	0.98(0)
AA2	Shifting cultivation	21	0.07(0.08)	0.77(0.32)
AA3	Fallow system cultivation	7	0.09(0.09)	0.90(0.12)
AA4	Ley system cultivation	33	0.12(0.12)	0.83(0.27)
AA5	Wetland rice cultivation	4	0.35(0.45)	0.53(0.43)
AA6	Irrigated cultivation (no rice)	70	0.11(0.10)	0.75(0.28)
AP	Perennial field cropping	3	0.13(0.13)	0.87(0.13)
AT	Tree and shrub cultivation	1	0.07(0)	0.98(0)
AT1	Non-irrigated tree crop cultivation	5	0.08(0.05)	0.72(0.29)
M	Mixed farming	4	0.08(0.04)	0.91(0.05)
MP	Agro-pastoralism	17	0.09(0.07)	0.69(0.35)
KEY	Drainage	Count	RMSE(sd)	R^2(sd)
V	Very poorly drained	22	0.09(0.06)	0.84(0.22)
P	Poorly drained	34	0.17(0.21)	0.72(0.31)
I	Somewhat poorly drained	26	0.11(0.09)	0.71(0.34)
M	Moderately well drained	75	0.12(0.10)	0.77(0.25)
W	Well drained	228	0.10(0.10)	0.76(0.28)
S	Somewhat excessively drained	12	0.13(0.09)	0.84(0.14)
E	Excessively drained	7	0.16(0.19)	0.78(0.27)
N/A	Not available	28	0.11(0.14)	0.71(0.32)
	Altitude	Count	RMSE(sd)	R^2(sd)
	<100m	106	0.12(0.15)	0.79(0.29)
	100-200m	33	0.14(0.12)	0.70(0.32)
	200-300m	43	0.09(0.08)	0.77(0.27)
	300-400m	12	0.07(0.05)	0.75(0.24)
	400-500m	16	0.10(0.06)	0.79(0.19)
	500-1000m	64	0.10(0.12)	0.74(0.28)
	1000-1500m	46	0.12(0.11)	0.77(0.29)
	1500-2000m	13	0.10(0.08)	0.62(0.31)
	>2000m	7	0.11(0.08)	0.69(0.31)
	N/A	92	0.12(0.10)	0.76(0.26)

Several typical examples were shown in Fig. A.6 to indicate the both the good and poor fitting of sigmoid function in the global dataset. With lime application, some profiles presented a decreasing trend, an opposite to the trend in local dataset. This trend was automatically captured and modelled using the sigmoid model. This increased the flexibility of the sigmoid model when fitting soil pH with either increasing or decreasing trends with depth in profile of the global dataset. Soils classes obtained acceptable fitting results, with RMSE values lower than 0.25 and R^2 values exceeding

0.50. Gypsisols and Kastanozems had the best fitting results with RMSE close to zero and R^2 close to 1. These soils, mainly found in the arid or semiarid region, had relatively high pH values (about 8) at the surface and a slightly increasing trend with depth due to high calcium accumulation (Vela et al., 2005). However, Plinthosols had a relatively worse fitting with RMSE of 0.22 and R^2 of 0.50. The pH values of Plinthosols in the global dataset ranged from 4.1 to 6 and can be classified as very strongly acidic to moderately acidic. The low pH values reflected the characteristics of Plinthosols, iron-rich and highly weathered soil (de Lucena et al., 2014). The hardpan formation due to fluctuating drying and wetting (Fritsch et al., 2007) might have caused the discontinuous and non-monotonic change of pH in those soil profiles and the sigmoid model could not account for that. Podzols also showed worse fitting results with RMSE of 0.16 and R^2 of 0.56. Sandy and acidic soil is typical feature of Podzols. Additionally, presence of Ae horizon with less Al and Fe was also common in Podzols. These resulted in the slightly peak feature of pH distribution, similar to the dataset shown in Bryk (2016).

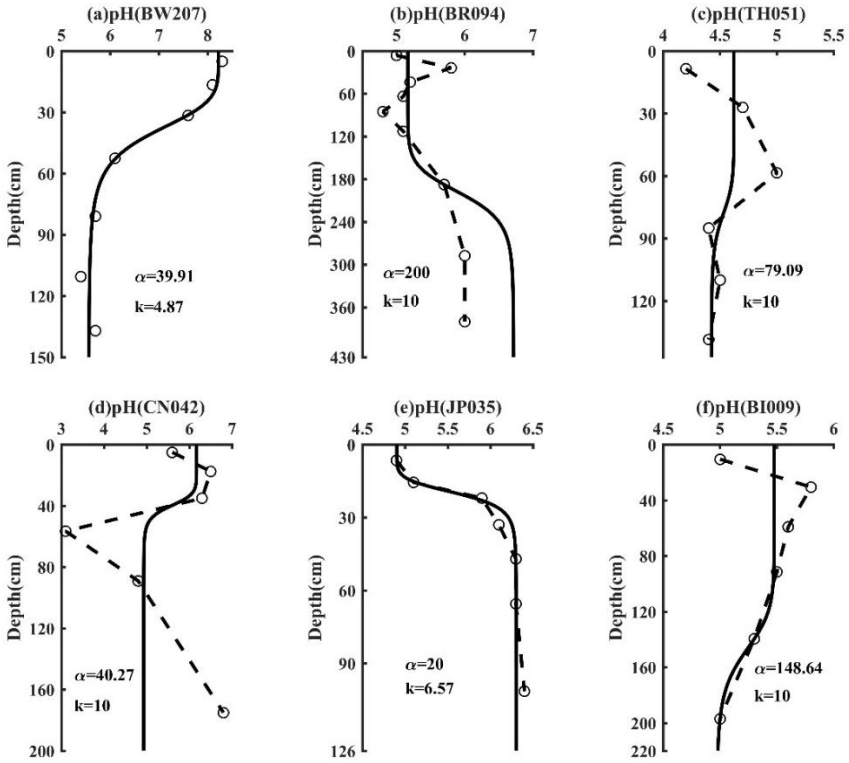

Fig. A.6. Examples of fitted sigmoid model in global dataset. (a) A decreasing trend of soil pH with depths; (b) Plinthosols; (c) Podzols; (d) profile under wetland rice cultivation; (e) Very poorly drained soil profile; (f) profile with elevation above 2000 meters. Dashed lines indicated linear lines connecting all the data points, and solid lines indicated fitting of sigmoid model.

Good fitting results with RMSE lower than 0.15 and R^2 above 0.70 were observed for the majority of the land uses except for wetland rice cultivation with observed RMSE = 0.35 and R^2 = 0.53. The soil pH values of Thai paddy soils of the global dataset were relatively low mainly due to the presence of acid sulfate soils and relatively acid soils in Bangkok Plain (Kawaguchi and Kyuma, 1974). In addition, soil pH of wetland rice cultivation system is highly influenced by fertilizer application which can significantly increase the soil pH (Li et al., 2009). Therefore, the combined effects resulted in a highly changeable profile condition and was not appropriately modeled by the sigmoid function. As for the drainage conditions, more than half of the profiles were well drained and showed moderate fitting performance (RMSE = 0.10, and R^2 = 0.76). The best fitting result was achieved for very poorly drained soil profiles with RMSE of 0.09 and R^2 of 0.84. The typical characteristics of very poorly drained soils that mainly includes Gleysolic and Organic soil according to the Soil Drainage Class (2013) were very similar as the local dataset. The soil profiles above 1500 meters exhibited comparatively poor fitting performance than the soil profiles below 1500 meters. However, Decker and Boerner (2003) and Bromley (1995) reported that elevation had little influence on soil pH values. We also think that the slight poor performance may not be due to the altitude and could be something else and needs more exploration.

Soil displays various features that are inherent to the factors and processes of soil formation. In turn, information on various factors and processes act as good indicators of various changes of soil properties. While most soil profiles in the global dataset with monotonic trends exhibited good fitting performance, specific soil classes, land use, and drainage factors obtained comparatively poor performance. The pH values of these profiles are usually highly changeable and may not exhibit a monotonic trend as a result of specific or combined effects of environmental or management factors. For example, the soil profiles from Plinthosols and wetland rice cultivation showed minima-maxima distributions which could be properly represented by 3rd order polynomial function. The soil profiles from Podzols and high elevation showed peak distributions. More site-specific depth function should be searched for the future work to fit such profiles.

A.3.4 Comparison of depth functions

The fitting performance of the sigmoid function in both local and global dataset was compared with that of the 3rd order polynomial function and the EAQSF with λ values of 0.001, 0.01, 0.1, 1, and 10. The predictive accuracy of depth functions was expressed by mean values of R^2 with standard deviation and RMSE with standard deviation and was presented in Table A.2. Profile D2

from the local dataset as an example was plotted with three depth functions to visually compare the vertical distributions in Fig. A.7. The 10λ EAQSF showed the weakest fitting performance in both local and global datasets with R^2 of 0.64 and 0.59 and RMSE of 0.30 and 0.20, respectively. Larger λ values increased the roughness of the spline function and reduced the accuracy (Bishop et al., 1999). The EAQSFs with λ value of 0.001, 0.01, and 0.1 were general among the best with R^2 close to 1 and RMSE close to 0 in both local and global datasets. The flexibility of EAQSF makes the fitting lines almost across all the points (Fig. A.7) and resulted in the highest accuracy outperforming other depth functions (Bishop et al., 1999; Liu et al., 2013; Taghizadeh-Mehrjardi, 2016). The standard deviations were highly correlated with the accuracy as the larger R^2 and smaller RMSE were aligned with smaller standard deviation (Table A.2).

Table A.2 Fitting results of three different depth functions

		Sigmoid	Polynomial	Splines				
				0.001 λ	0.01 λ	0.1 λ	1 λ	10 λ
Local	RMSE	0.08	0.10	0.02	0.01	0.01	0.09	0.30
	(sd)	(0.08)	(0.09)	(0.01)	(0.01)	(0.01)	(0.06)	(0.11)
	R^2	0.97	0.95	0.998	0.999	0.999	0.96	0.64
	(sd)	(0.05)	(0.06)	(0.002)	(0.001)	(0.001)	(0.03)	(0.11)
Global	RMSE	0.11	0.08	0.03	0.03	0.02	0.07	0.20
	(sd)	(0.12)	(0.09)	(0.03)	(0.03)	(0.01)	(0.06)	(0.13)
	R^2	0.76	0.87	0.98	0.98	0.99	0.93	0.59
	(sd)	(0.29)	(0.18)	(0.02)	(0.02)	(0.01)	(0.07)	(0.19)

The sigmoid model, the 3rd order polynomial function, and the 1λ EAQSF achieved comparable and reasonably well fitting results, but the performance sequence was different for the local and the global dataset. In the local dataset, these functions provided good fitting results of R^2 values above 0.95 and RMSE values lower than 0.10. The 3rd order polynomial function showed a non-monotonic trend in deep soil layers (Fig. A.7), resulting in a slightly worse fitting. However, in the global dataset, 3rd order polynomial function with R^2 of 0.87 and RMSE of 0.08 outperformed the sigmoid model with R^2 of 0.76 and RMSE of 0.11. This indicated that the vertical distribution of soil pH are not always monotonic increasing or decreasing around the world. This also supported the previous discussion that complex profile conditions and non-monotonic distributions e.g. peak distribution and minima-maxima distribution existed worldwide as a result of various soil-forming factors and processes (Fig. A.6). These distributions should be modeled by more soil type-specific and flexible functions in the future. 3rd order polynomial function with more changeable trend can

serve as a substitute for the sigmoid model in some situations.

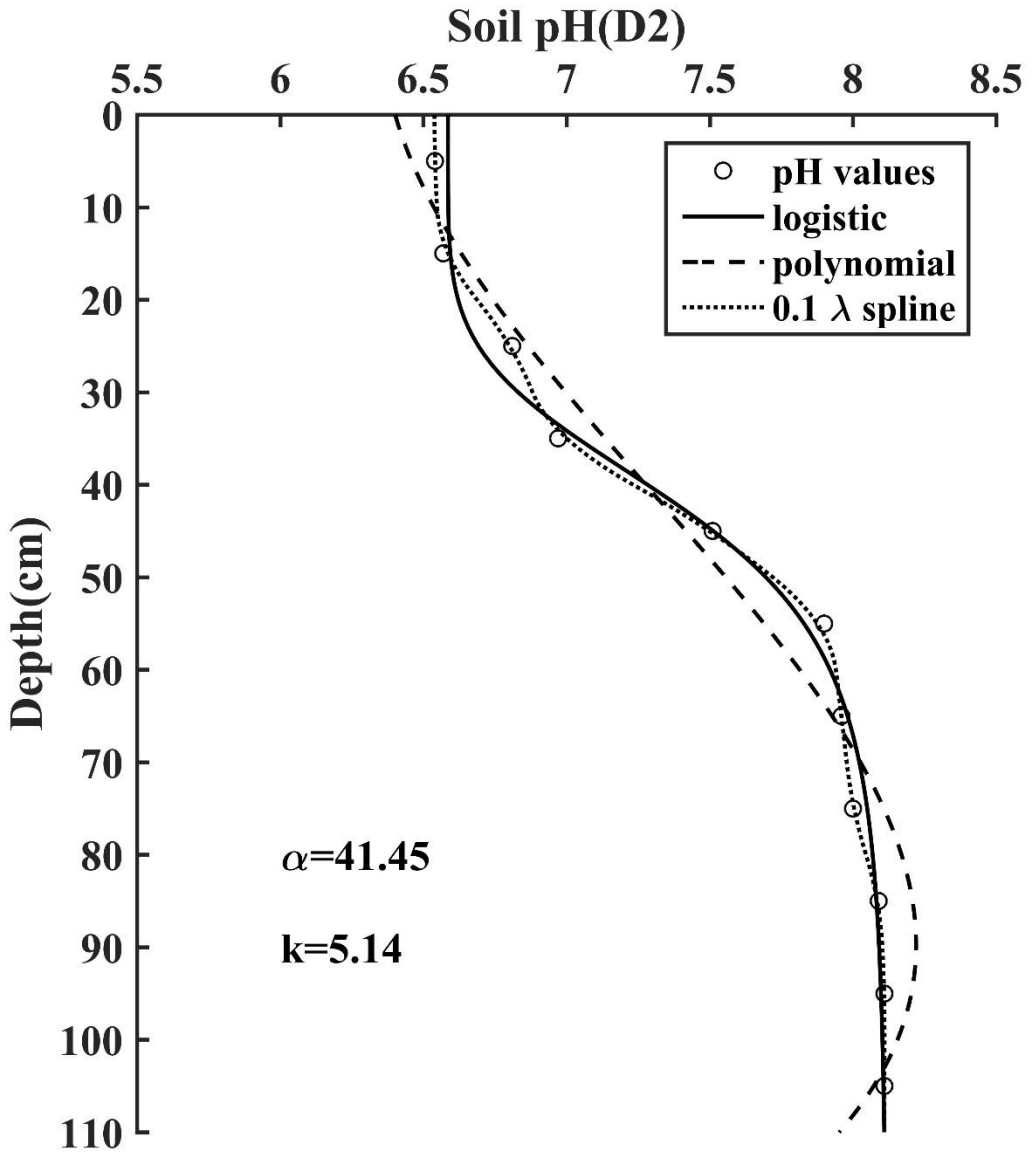

Fig. A.7. Comparison of vertical distributions of fitted sigmoid model, 3rd order polynomial function, and 0.1λ EAQSF (profile D2 of the local dataset).

Accuracy is not always the only criterion for the best model. Some other factors such as interpretability, simplicity, generality, the number of parameters, and the mass balance issue should also be compared to determine an efficient profile depth function (Table A.3). It is noteworthy that the sigmoid function has four parameters and all of these parameters have a clear explanation reflecting pH distribution in soil profiles and physical conditions of agricultural fields. Initial values of parameters need to be carefully chosen, as different values lead to different optimization procedures with the undesired locally optimal solution (Webster and Lark, 2013). With a clear

physical explanation, initial values of four parameters can be easily chosen for a sigmoid function, resulting in a more reasonable solution. In contrast, four parameters in 3rd order polynomial function lacks physical explanations and the uncertainties and randomness in choosing the initial parameter values might not escape from locally optimal solutions. In addition, determining the degrees of polynomial functions is generally arbitrary and the variation could always be fitted by higher level polynomials (Webster, 1978). However, 3rd order polynomial function showed advantages in fitting the minima-maxima pattern in many profiles of global dataset. The EAQSF showed the highest flexibility and accuracy in fitting variation of any soil properties. However, it is not unique to specific soil properties and could not represent the natural pedological feature of that specific soil property. It's distribution trend changed for individual soil profile and lacks generality. Moreover, EAQSF has the ability to keep mass balance which means the area under the fitting curve is the same as the total area of soil horizons. By comparison, sigmoid function and polynomial functions cannot guarantee a mass balance, while a better fitting (lower RMSE and higher R^2) obtained by these functions better keeps the mass balance (Table A.3).

Table A.3 Comparison of different depth functions

Factors	Sigmoid	Polynomial	Spline
Interpretability	+++	+	+
Simplicity	+++	++	+
Accuracy	++	++	+++
generality	+++	++	+
Parameters	4	4	more
Mass balance	+	+	+++

A.4 Conclusions

Vertical distribution of soil pH usually displays such a pattern of uniform soil pH in the plow layers due to the mixing effects of tillage and in bottom layers due to the less disturbance, and a transition zone in between. A closed form sigmoid model with four parameters was proposed to quantify the vertical distribution of soil pH in agricultural fields. The developed model can be used to predict pH values of any location in depth and prepare for 3D digital soil map. This model obtained good fitting performance in a local dataset and reasonably moderate performance for the majority of profiles in a more complex and changeable global dataset. However, more type-specific depth functions are also needed to take into account all the possible factors in the future work. Comparisons of sigmoid models, 3rd order polynomial functions and EAQSFs showed that sigmoid

model was superior with consideration of interpretability, simplicity, and generality. EAQSF provided the most flexible fitting and accurate results and showed advantages in keeping mass balance. However, it was more complex to implement and lacked physical explanation for specific soil properties and fixed trends. The 3rd order polynomial function was inferior considering all the factors, while its unique feature which displayed non-monotonic distribution at the bottom layers can be further explored to fit more complex distribution of soil properties e.g. minima-maxima distribution.

However, the accuracy obtained in this study is regarded as the internal accuracy, which is always overly optimistic (Chatfield, 1995). An independent dataset (the soil pH at other depths within the profile) should be used to test the model, and this is known as external accuracy (Williams, 1996). In addition, the samples used in this study were collected at 10 cm depths, which means these are average values instead of real values of specific depths. Therefore, for future work, in-situ soil pH measurement which can instantly measure soil pH at 1 cm intervals can be used to develop a more reliable model and test the external accuracy of the model (Matthiesen, 2004). The sigmoid function was proposed and tested specifically for soil profiles in agricultural fields, its application in other land use types might be restricted. For example, the soil profiles descriptions in forest soil (Joshi and Negi, 2015; Krueger et al., 2016) and wetland soil (Lin et al., 2015) were too irregular to be modeled by a simple sigmoid function.