Development of operational methods to predict soil classes and properties in Canada using machine learning

by

Xiaoyuan Geng

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in

The Geography and Environmental Studies

Carleton University
Ottawa, Ontario

Abstract

With a limited amount of arable land, worsening soil degradation, and plateauing of crop yields, the maintenance of limited soil resources to guarantee agriculturally productive soils and to support essential ecosystem services is critical in Canada and globally. The main objective of this thesis is to study and develop practical methods to cost-effectively renew soil and soil landscape information in Canada. Given the limited amount of point-based soil data in Canada, machine learning-based predictive methods were studied for possible operational use. Machine learning methods require point soil data both for training purposes and validating the predictions. When soil data points are lacking, pseudo-point soil data are mined through soil survey polygon maps. In places where detailed soil surveys exist for soil class mapping, fully randomized pseudo-soil point data mining with random forest-based machine learning achieved a prediction accuracy as high as 74%. The prediction was further improved to a high of 78% by using an ensemble of multiple machine learners. Soil properties such as bulk density can be predicted either directly using point soil data or indirectly via predicted soil classes which are associated with reported soil property values. In this study, sampled soil bulk density values were used to predict soil bulk density across the study watershed. The predicted bulk density values come with uncertainty ranges, computed using residual kriging. Soil class prediction (and soil bulk density) may be carried out using environmental covariates from different sources. It is shown that where surficial geological material data are lacking, time series microwave remotely-sensed data, specifically Sentinel-1A synthetic aperture radar (SAR) imagery, can be used to delineate soil spatial patterns which are hypothesized to be linked to the spatial distribution of surficial geological materials. Through this study, a cost-effective

work flow and solutions for predictive soil mapping needs in Canada were developed for operational use.

Acknowledgements

A person's success is often the combination of talent, hard-work and luck. Completion of this work is not much the success rather than a milestone of my life! With a lot of hard work, I have had much more luck than talent. The luck that I have had is not the kind of winning a lottery. It is the kind of having many loving and supportive people through my life and especially in this academic endeavor: my forward thinking parents who supported and kept me in schools even during very difficult circumstances; my loving and supportive family which has always been my comfort and inspiration; my great colleagues who allowed me to step on their shoulders, who mentored me and most importantly helped me! Special thanks go to Drs. Juanxia He, Yefang Jiang and Bert VandenBygaart for their help! Thanks Drs. Scott Mitchell, Elyn Humphreys, Murray Richardson, Xulin Guo, and Richard Amos for their guidance and critical review of this work.

For my late father whose last word to me was to go this far!

Table of Content

Abbreviation

AAFC, Agriculture and Agri-Food Canada
ALOS, Advanced Land Observation Satellite
ANN, Artificial Neural Network
ARY, Alberry
ASTER, Advanced Space-borne Thermal Emission and Reflection Radiometer
BD, Bulk Density
BMPs, Beneficial Management Practices
BN, Bayesian Network
CanSIS, Canadian Soil Information Service
CART, Classification And Regression Tree
CBR, Case Based Reasoning
CFS, Canadian Forest Service
CHAID, CHi-square Automatic Interaction Detector
cLHS, conditioned Latin Hypercube Sampling
CNN, Convolution Neural Networks
CPD, Crapaud
CTW, Charlottetown
DEM, Digital Elevation Model
DSM, Digital Soil Mapping
DSMART, Disaggregation and harmonization of Soil MAp units through Resampled classification Trees
EW, Sentinel-1A Extra Wide swath
FCM, Fuzzy C-Means
FFNN, Feed-Forward Neural Network
FMF, Fuzzy Membership Function
GAM, Generalized Additive Model
GDA, General Discriminant function Analysis
GDFA, General Discriminant Function Analysis
GIS, Geographic Information System
GLM, General Linear Model
GRD, Ground Range Detected
GRM, General Regression Model
GUI, Graphic User Interface
ID3, Interactive Dichotomizer 3
LHS, Latin Hypercube Sampling
LIDAR, Light Detection and Ranging
LOOCV, Leave One Out Cross Validation
MF, Membership Functions
NBMI, Normalized Radar Backscatter (Soil) Moisture Index
NDVI, Normalized Difference Vegetation Index
NSDB, National Soil DataBase
OK, Ordinary Kriging
PALSAR, Phased Array type L-band (15-30cm wavelength) Synthetic Aperture Radar
PEI, Prince Edward Island

PFRA, Prairie Farm Rehabilitation Administration
PSM, Predictive Soil Mapping
QUEST, Quick, Unbiased, Efficient Statistical Trees
RBR, Rule-Based Reasoning
RCM, Radar Constellation Mission
RF, Random Forest
RK, Regression Kriging
SAGA. System for Automated Geoscientific Analyses
SAR, Synthetic Aperture Radar
CLORPT, Climate, Organism, Relief, Parent material, and Time
SI, Semantic Import model
SIE, Soil Inference Engine
SLC, Soil Landscapes of Canada
SOC, Soil Organic Carbon
SOFM, Self Organizing Feature Map
SOLIM, SOiL Inference Model
SRTM, Shuttle Radar Topographic Mission
SVM, Support Vector Machine
SWAT, Soil Water Assessment Tool
TPI, Topographic Index
TWI, Topographic Wetness Index
USDA, United States Department of Agriculture
MRVBF, Multi-Resolution Valley Bottom Flatness
WSO, Winsloe
WRB, World Reference Base
ZSC, Stream Complex

List of Tables

List of Figures

Chapter 1: **Toward predictive soil mapping in Canada**

1.1 Introduction

Over the last two decades, traditional soil mapping has been gradually replaced with statistically based and data-driven predictive soil mapping (PSM) approaches. The investment in predictive soil mapping has been mainly centered on research and development of predictive soil mapping methodologies (Scull et al, 2003; Hartemink and McBratney, 2008; Sanchez et al., 2010; Minasny and McBratney, 2016; Hengl et al., 2017; Rossiter, 2018; Zhang et al., 2017; Hengl and MacMillan, 2019). However spatially explicit soil information at various scales (e.g. field, watershed, region, nation and world) is urgently needed to address sustainability, changing environment and ecosystem services and related issues (Keskin and Grunwald, 2018). With disappearing investment in conventional soil surveys in Canada, an operational PSM framework and methods are needed to fill the void of missing up-to-date soil data and information at landscape scales. The overall goal of this thesis is to research options, and to develop an adaptive operational framework, for PSM in the context of Canadian landscapes.

1.2 Thesis objectives

This thesis has three main objectives. The first aims to study the use of expert knowledge that is embedded within legacy soil surveys and reports using machine learning techniques to predict soil classes and soil properties. The second is to investigate geospatial sampling of both legacy survey coverage and new field data to identify an optimized, statistically-sound and cost-effective sampling method, which is essential for operational predictive soil

mapping. The third is to study adaptive ways of using environmental covariates from different sources and at various scales or resolutions for PSM. When topographic and surficial geological material features cannot be effectively used for PSM, remotely-sensed time series information including Synthetic Aperture Radar (SAR) microwave technology may have great potential.

Mining knowledge from data sources at various scales and vintage soil survey databases is challenging. For the areas where legacy soil point or pedon data and fine soil survey maps (e.g. 1:20,000 scale soil survey) exist, machine learning training information can be mined and used under certain assumptions. Soil survey area class maps often have a complex data structure and higher level of generalization. For example, a soil component in the national soil database in Canada means a soil type has been mapped and described in a map unit or soil polygon. A mapped soil polygon can have more than one soil component but not include the explicit spatial locations of the soil components. A soil map polygon with a single component claims that only one major soil exists in that mapped area. However, extracting knowledge from even a single component soil polygon would require the assumption of a homogenous soil polygon.

When using soil genesis model-based knowledge to infer soils across landscapes, the scale or resolutions of the environmental covariates are often subjectively selected, processed and represented as users' preferences. The importance and influence of environmental covariates or features on PSM needs to be better understood.

For areas where landscapes are generally flat, other easy-to-observe soil covariates such as topographical features become less important. However, the spatial patterns of soils along with the underlying parent materials still exist. Remotely-sensed data are expected to reflect soil spatial patterns in these cases. The use of optical remote sensing time series data for PSM has been studied in Canada as well as in other regions (Liu et al., 2012). Zhu et al. (2010) used remote-sensing with a response surface hypothesis, asserting that locations with different soil conditions would have different feedbacks after major rainfall events, and the different feedbacks would be detectable by the differences in remotely-sensed signatures. In this study, the response surfaces of time series remotely-sensed data, specifically microwave radar data before and after a major rainfall event, are hypothesized to be spatially correlated with soil type / soil parent materials and associated properties. In this manner, the response surface framework assumption will be tested along to help evaluate the usefulness of microwave remote sensing data for PSM of soil properties.

This thesis is presented in seven chapters. Following this brief introduction which outlines the major objectives of the thesis, Chapter 2 presents background information on Canadian soil surveying and recent progress in PSM, especially in Canada. Chapter 3 details further background information on several frequently- used inference algorithms in PSM. It then introduces the core algorithms of tree-based random forest, fuzzy logic and neural network principles, and summarizes some Canadian and international cases of those inference methods (to further contribute to understanding of these inference algorithms, specific examples are given in the appendices). The main objective of Chapter 4 is to develop a method for soil type inference based on legacy soil survey data, validated using new field

survey/inspection. Chapter 5 presents two key methods of soil property mapping. One is based on the inferenced soil type map and another one is based on newly sampled point data. Pros and cons of the two methods are discussed. Chapter 6 examines the use of time series microwave radar data for PSM with a focus on soil parent materials as an important soil forming factor. Chapter 7 offers a summary and discussion of a practical PSM framework in the context of the state of soil data and business needs in Canada. Future research directions are outlined.

Chapter 2: **From conventional soil survey to predictive soil mapping: a review**

Due to diminished conventional soil survey activity in Canada, a predictive soil mapping approach is the only cost-effective way to produce up-to-date soil and soil landscape data. However there is a lot of value in legacy data, and this research was designed to carry as much legacy knowledge and information embedded in conventional soil surveys as possible. Much of the background to this work included collaborative efforts such as Geng et al. (2010b), Schut et al. (2011), Nyiraneza et al.(2017), and Geng et al.(2016). The specific research questions addressed in this thesis build upon that prior work, and contribute back to the author's broader collaborations in the global soil mapping community.

2.1 Canadian soil survey and the national soil database

Canada has a long history of soil surveys, dating back to 1914. These surveys provided information to address real needs and issues of the time, including soil and water conservation, the improvement of farm productivity and management, land reclamation, and land use planning. Soil surveys provided key knowledge for adapting to prairie droughts and provided baseline information for major national and provincial initiatives including the creation of the Prairie Farm Rehabilitation Administration (PFRA) in 1935 and the Agricultural Rehabilitation and Development Act (ARDA) in 1961. By the 1970s and 1980s, there were field inventories and associated maps and reports for most populated areas of Canada (McKeague and Stobbe 1978; Anderson and Smith, 2011). Soil maps were

generally published at scales between 1:20,000 and 1:250,000 for regional land management and planning purposes. Government-led soil surveys have declined in recent years, as well as the extension capacity formerly provided by the Prairie Farm Rehabilitation Administration, which closed in 2010. Fortunately a legacy of maps and data remain today under the stewardship of the Canadian Soil Information Service (CanSIS), where soil data are distributed via an array of on-line distribution mechanisms including the website at http://sis.agr.gc.ca/cansis/ (Accessed May 7, 2020).

Beginning in 1972, AAFC developed a custom Geographic Information System (GIS) to process and store soil maps and data in electronic format. Many of the world's first GIS were also being developed during that time period, also using custom software. Since then the technology used to support the system has continuously evolved, but the function of the system has remained focused on making soil information (databases, maps, reports and manuals) readily accessible to users. CanSIS publishes various scales of soil (interpretive) maps via the website (http://sis.agr.gc.ca/cansis, accessed May 7, 2020). Detailed soil maps exist for approximately 2% of Canada's total land area, consisting mainly of the agricultural areas in each Canadian province or territory (Figure 2-1).

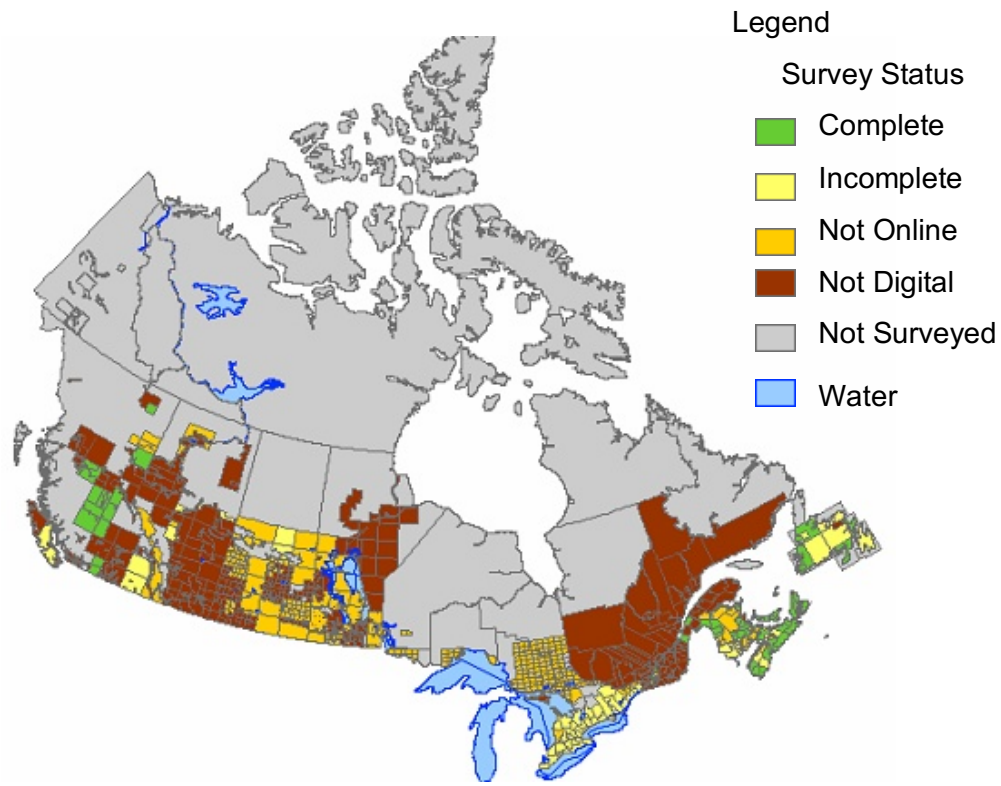


Figure 2-1 Coverage of detailed/semi-detailed soil surveys in Canada (Geng et al., 2010b, data from http://sis.agr.gc.ca/cansis/nsdb/dss/v3/index.html, accessed May 7, 2020).

Over the last decade, AAFC, with provincial partners, has attempted to create seamless provincial digital soil maps at a standard scale to replace the many individual detailed map sheets produced at a range of scales covering local, county or rural municipal jurisdictions. For these new seamless coverages, soil naming conventions and the attribute data structures are correlated and standardized for soil units/regions (polygons) across the country.

Soil Landscapes of Canada (SLC) maps are generalized maps published at 1:1,000,000 scale. Several versions of the SLC maps have been published over the last 30 years reflecting changes in data availability, management and technology. SLC version 2.2

covers all of Canada as shown by the polygon outlines in Figure 2-2.

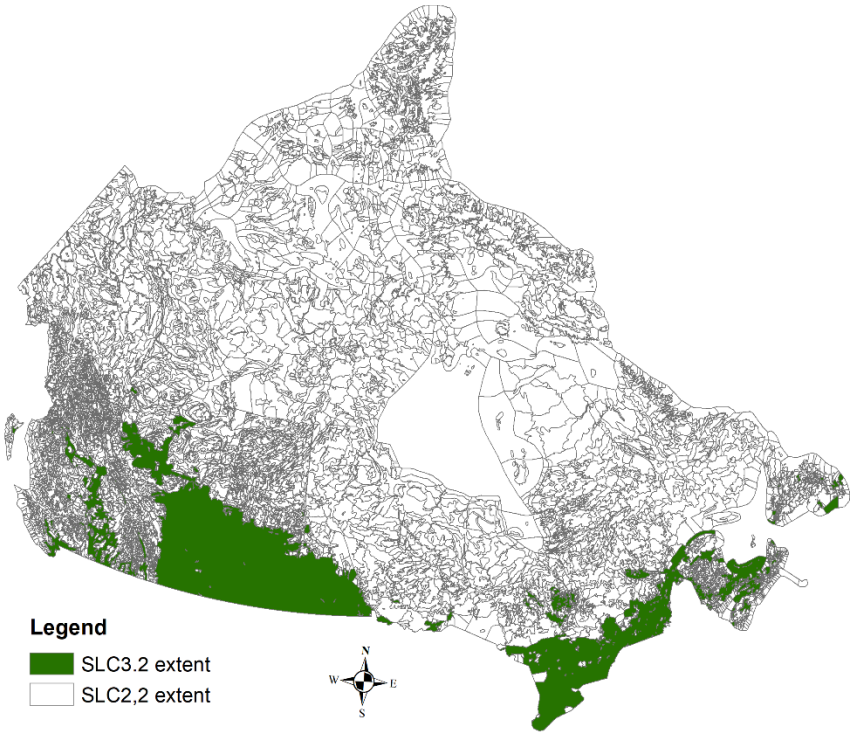


Figure 2-2 Coverage (in green) of the latest version of the Soil Landscapes of Canada (data source http://sis.agr.gc.ca/cansis/index.html, accessed May 7, 2020)

Within the agricultural extent of the country, mapping has been upgraded and published as the SLC version 3 series. These versions as well as various detailed map products, reports and technical manuals are distributed freely via the CanSIS website (Schut et al., 2011). CanSIS also holds about 7,000 pedon records from across Canada. A standard pedon "is the smallest, three-dimensional unit at the surface of earth that is considered as a soil", and it usually has a surface area of approximately 1 m^2 (Soil Classification Working Group, 1998). The number of harmonized pedon data records is increasing gradually as part of ongoing efforts of the national pedon database expansion work in CanSIS. However, most of the sampled pedons are located in the southern regions of Canada where agricultural

land predominates (Figure 2-3).

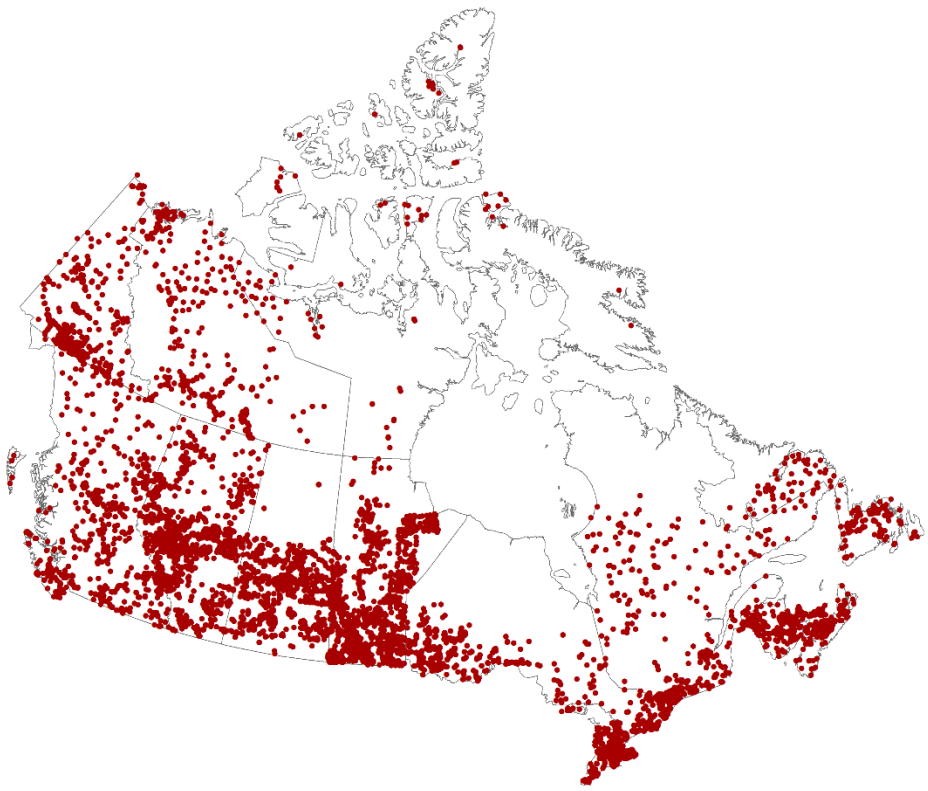


Figure 2-3 Coverage of the harmonized national soil point or pedon data (revised from Geng et al., 2010b)

The locations of the legacy pedons were often selected based on ease-of-access and access permission, therefore the distribution of those pedons is skewed and clustered. Many additional pedon records are held outside of CanSIS by other agencies and organizations such as the Canadian Forest Service (CFS). CFS recently released about 2,000 pedon data points, mainly from forested areas across Canada (Shaw et al., 2018). Pedon data from other institutions are not readily accessible. The usability of all the pedon records has yet to be fully assessed in terms of location accuracy, attribute completeness and laboratory methods of measurements. Pedon data from different sources were often measured with

inconsistent field and laboratory methods. Since the legacy soil pedons were not sampled with statistical sampling designs, the amount of points and their distribution may not be statistically representative for predictive soil mapping. Also, data collected over many decadal time spans may not be appropriately combined in some cases, since many soil properties change considerably over time (Judith et al., 2017).

2.2 Soil mapping procedures and legacy soil data structure

Although soil develops over a continuum of changing topography and underlying surficial geological materials, soils in Canada have been classified, modeled and mapped discretely based on the relationship between soil and soil forming factors, namely climate, topography, parent material, organisms and time (Jenny, 1994), described in further detail in Section 2.3. Conventional soil survey or mapping in Canada has been characterized by area class maps and reports including associated soil attribute data. In the last hundred years, soil surveys in Canada have focused on the identification of bodies of related soils that can be recognized as natural units, and their delineation on maps using conceptual models and tacit knowledge held by soil surveyors and pedologists (Coen, 1987). Major soil types have been identified, described and delineated on a map base along with direct field observations, supplemented by indirect inferences based on aerial photo interpretations. Although soil surveys in Canada have been created following similar methods used by many other countries such as the former Soviet Union, they share similar limitations: the identification and delineation of soil bodies was the product of pedological concepts and models developed by individual mappers, which were often not replicable (Brevik et al., 2016).

The main steps usually followed in a soil survey in Canada are (Geng et al., 2010b):

(1) soils observed within a survey area are grouped into a limited number of soil names (series or associations) based on properties that are relevant to the survey objectives;

(2) a field pedon investigation samples and describes each major named soil and the samples are submitted for laboratory analyses, to determine soil attributes;

(3) soil map polygons are delineated and attributed to describe the portions of the landscape associated with each soil name; and,

(4) each named soil is identified with a unique code (e.g. Province Code + Soil_Code + Modifier + Land_Use), thereafter used to reference relational soil information within the National Soil DataBase (NSDB).

The data model used to manage Canadian soil data has evolved to meet soil resource data business needs and changing technological capacity. The current data model of the NSDB within CanSIS is illustrated in Figure 2-4. This relational data model starts with polygon feature vector GIS data. SLC geometry (Figure 2-4) is presented at 1:1 million scale. The scales of the Detailed Soil Survey range from 1:10,000 to 1:250,000 (Figure 2-4).

Pedon data including both point feature class and tabular field and laboratory information are the foundation of the soil layer table and higher-level data derivations in the NSDB. The layer table is constructed for each soil name with attributes calculated by averaging values contained in all pedons representing that name. Often there are not enough sampled pedons for true statistical means. Indeed some of the named soils have never been properly

sampled. In cases when no analytical pedon data exist for a soil name, layer table attributes are estimated or derived by using pedo-transfer functions (Van Looy et al., 2017). The layer tables are used to support both detailed and SLC mapping and accessed by any algorithm or scientific model run against the NSDB.

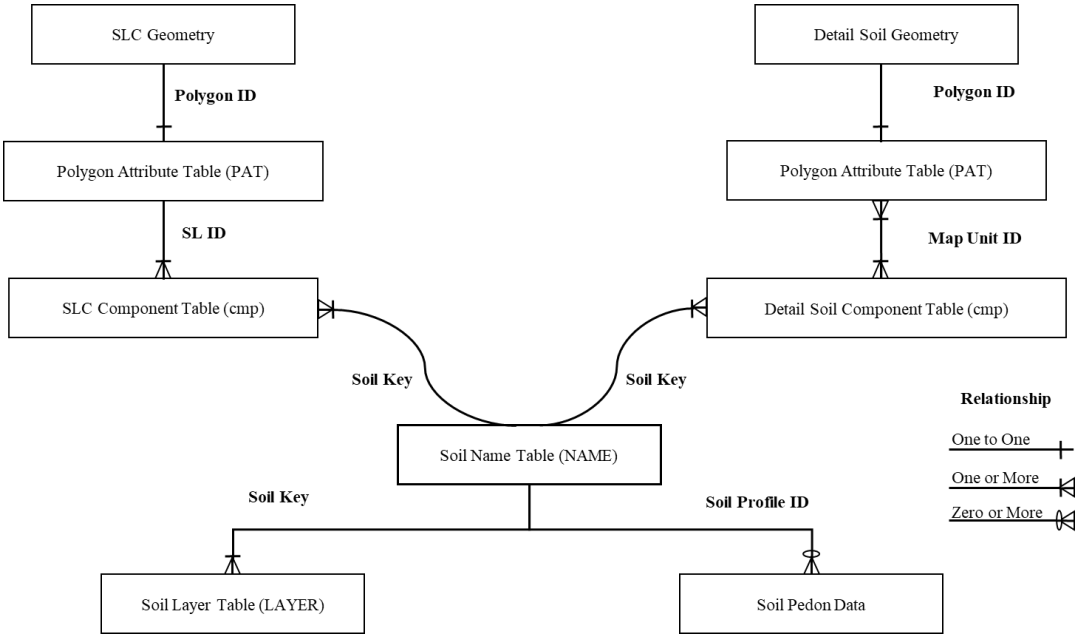


Figure 2-4 Relational data model of the National Soil Database in Canada (updated from Geng et al., 2010b; see also http://sis.agr.gc.ca/cansis/nsdb/index.html).

2.3 Soil genesis and predictive soil mapping

Soil forms on a continuum of topography and surficial material and as such, spatial variability of soils is inherent in models describing soil genesis and development. In practice, soil surveys are typically characterized by treating soils as geographic bodies which are closely related to the soil forming factors (Glinka, 1927). Jenny (1941) conducted extensive experimental validation on the theory of soil development and soil–

environment relationship across the landscapes in USA. Since then, the recognition of the soil and environment relationship framework has been the foundation of institutional soil survey and mapping in most parts of the world including Canada. Hudson (1992) further addressed that this soil-environment framework is the paradigm which is the foundation of conventional soil survey and mapping.

With this scientific paradigm, soil scientists and soil surveyors model the relationship between soil (classes and properties) and environmental variables, namely climate, topography, parent material, organisms and time, using a state factor equation called CLORPT (CL = climate condition at a point; O = organisms including land cover; R = relief or topographic attributes; P = parent or surficial geological material; T = time or age) and expressed as a theoretical soil-landscape relationship (i.e. Sc or a= $\int$(soil forming factors)dt) where Sc is soil class and Sa is soil attribute (Jenny, 1941). With increased human impacts on the environment including soils, Grunwald et al. (2011) and Thompson et al. (2012) have proposed to extend the conceptual relationship between soil and soil forming factors by including anthropogenic forcing, such as human-enhanced water erosion. The CLORPT model also doesn't account for covarying soil forming factors. To truly reflect the relationship between soil and soil forming factors, the universal model of spatial variation recognizes that both the deterministic and stochastic components of soil forming factors need to be modeled (Matheron, 1971). Such soil-environment relationship frameworks, especially the CLORPT model and the universal variation model, have been the foundation of recent predictive soil mapping (Scull et al, 2003; Hartemink and McBratney, 2008; Sanchez et al., 2010; Viscarra Rossel and Bui, 2016; Hengl et al., 2017;

Zhang et al., 2017).

Predictive soil mapping, also referred to as digital soil mapping, is the computer-assisted production of spatial data or maps of soil types and soil properties using the structured knowledge of soil and its relationship with environmental variables (Hudson, 1992; Bui, 2004; Shi et al., 2009). PSM involves the creation and population of spatially explicit soil information from field and laboratory observations coupled with spatial and non-spatial soil inference systems (Lagacherie and McBratney, 2007; Grunwald, 2009; Hengl and MacMillan, 2019). Soil classes and properties include categorical examples such as soil name, and interdependent continuous variables such as bulk density.

The interpolation, prediction and inference of soil classes and/or soil property distributions have been dealt with through diverse methods in the last twenty years (Moore et al., 1993; Zhu et al., 1996; McKenzie and Austin, 1993; McBratney et al., 2000, 2003; Kempen et al., 2009). Nussbaum et al. (2018) organized the methods of PSM into six groups: conventional statistically-based, such as linear regression models; geostatistically-based (e.g. kriging); generalized additive models (GAMs); binary classification and regression trees (CARTs); ensemble machine learning such as random forest (RF); and hybrid methods (e.g. model averaging over outputs of multiple algorithms). Cianfrani et al. (2018) broadly divided the frequently used PSM methods into two groups of geostatistical and predictive statistical approaches. A good example of the geostatistical approach is ordinary kriging (OK), which uses observation point data to interpolate information while factoring in spatial autocorrelation. The predictive statistical methods use soil-related covariate and

point soil data. From these recent reviews, predictive statistically-based approaches especially RF and hybrid ensembles have been more frequently used and have generally been found to be the most accurate (Cianfrani et al., 2018; Nussbaum et al., 2018). Zhang et al. (2017) singled out the knowledge-based approaches (especially fuzzy set and logic) from the others. Although the operational use of fuzzy logic-based methods are limited to small mapping areas with a limited number of soil classes, the use of the fuzzy set and fuzzy logic methods has been extended and applied successfully to purposive sampling designs, which is an important component of predictive soil mapping (Zhu et al. 2008; Yang et al., 2011). The fuzzy logic framework has guided digital soil mappers to effectively organize expert knowledge from legacy data and tacit knowledge.

The relationship described in the CLORPT model is nonlinear (Scull et al., 2003). The need for capable numerical solutions or statistical approaches is inevitable. In the context of predictive soil mapping using statistical approaches, there exists a multitude of methods that allow computers to learn and to infer. Machine learning methods can not only handle nonlinear relationships between soil and soil forming factors, but can also quantify relationships and predict soil type and property distributions across landscapes. From training to inference, machine learners also mimic the way traditional soil surveys were derived through mental models and the tacit knowledge of soil scientists or pedologists (Bui, 2004; Walter et al., 2007). In attempting to deal with the nonlinear relationship between soil class/property and environment variables, frequently used machine learners include random forest (Heung et al., 2017; Hengl et al., 2017), artificial neural networks (ANN) (Zhu, 2000; Behrens et al., 2005; Brus et al., 2008), support vector machine (SVM)

(Ballabio, 2009), fuzzy logic algorithms (Mazaheri et al., 1995; Zhu et al., 1996; McBratney and Odeh, 1997; De Gruijter et al., 1997; De Bruin and Stein, 1998; Zhu et al., 2010), and Bayesian statistical models (Beguin et al., 2017). Indeed, hybrid ways of using ensembles of many machine learners have shown operational potential (Cianfrani et al. 2018; Hengl and MacMillan, 2019).

2.4 Soil expert knowledge and data mining

From an artificial intelligence and expert system perspective (Rich, 1987), soil scientists and soil surveyors are seen as (tacit) knowledge experts. These soil professionals use available environmental information such as elevation, slope, aspect, vegetation, and parent material types to construct the relationship between soil and soil forming variables, and then infer or predict the soil classes and soil attributes, including behavior, within or beyond the areas where information and knowledge are derived.

People can partially acquire this tacit knowledge through classroom learning. Soil maps and legends also partially carry some structured knowledge (Bui, 2004). However, much of the expert knowledge of soil and its environmental relationships (including behavior) are acquired through field work and professional training. How to retain and apply (tacit) expert knowledge is a challenge. Hudson (1992) suggested that this could be dealt with by advancing pedological teaching and linguistic representation of the tacit knowledge. Tacit or expert knowledge may be mined using proper knowledge acquisition methods such as structured or unstructured interviews during expert system implementations (Zhu et al., 1996; Qi et al., 2006).

The expert knowledge used to produce soil maps is often characterized in scale and space. The scale refers to the geographical (e.g. global vs local) coverage of the knowledge: some knowledge is more general or global and can be applied to a broader mapping area; some is more specific and local and can only be applied for a limited mapping area. Shi et al., (2004, 2009) used the term "rule-based reasoning" (RBR) when they conducted soil inference using global knowledge and "case-based reasoning" (CBR) using local knowledge. Rule-based knowledge represents general or global knowledge between soil and soil forming factors. For example, a general rule can be 'soil A is found in places where the local slope gradient is less than 3%'. Rule-based knowledge is often constructed by assuming that soil forming variables are independent of each other since a soil expert normally works with one environmental variable at a time when he/she is explicitly defining the relationship between soil and soil forming variables. Qi et al. (2006) borrowed a term 'prototype' from the expert system domain to describe RBR for a predictive soil mapping study. Alternatively, since the soil-environment relationship is not fully understood, it can be applied in the form of cases rather than general prototypes. Case-based knowledge can be used to infer soils across landscapes and even regions if the hypothesis that environmentally similar units tend to have similar soils (cases) is valid (Hudson, 1992).

Soil development knowledge is also buried in soil survey reports, maps, and attribution tables of digitized soil maps. Among various examples of soil survey information, soil point data are most valuable for both numerical modeling-based and knowledge-based

PSM (Adhikari et al., 2014) and its validation. When recent field point soil data are lacking or insufficient, legacy soil point data from soil profile databases or that derived by data mining procedures become vital. Legacy soil survey data mining has been utilized extensively in PSM (Dobos et al., 2010; Vaysse and Lagacherie, 2015; Bulmer et al., 2016; Heung et al., 2014, 2017). In Canada, where nationwide soil point data are sparse (Geng et al., 2010b), detailed soil surveys range in scales from 1:10,000 to 1:65,000 in many agricultural regions. Detailed soil surveys can be used to extract point soil information and other knowledge for PSM, and are often at the level of detail such that many mapped soil polygons contain only one soil type or component (Geng et al., 2016).

With recent advances in computing and geospatial science, data mining methods can facilitate the acquisition, representation and application of knowledge for PSM. For example, using GIS overlay and simple logic operations, pseudo-point data can be mined from legacy soil data along with related co-variables using the fuzzy c-mean algorithm on soil attributes and the environmental variables that influence soil (e.g. covariates), where the centroid of those classified patterns are referenced for pseudo point or purposive sampling design (Liu et al., 2012; Yang et al., 2010; Zhu et al., 2016). Data mining tools can include manual (Nauman et al., 2012), semi-automated (Thompson et al., 2010), and automated model-based methods, and include tools such as ID3 (Interactive Dichotomizer 3), C5 or See 5, CART (classification and regression tree), QUEST (Quick, Unbiased, Efficient Statistical Trees) (Breiman et al., 1984), CHAID (Chi-square Automatic Interaction Detector) (Kass, 1980) and CRIUSE (Behrens et al., 2010). Odgers et al. (2014) developed a model-based data mining tool named Disaggregation and harmonization of

Soil MAp units through Resampled classification Trees (DSMART) to mine and use soil knowledge from legacy soil survey data. DSMART with embedded C5-like algorithms is used to transform and disaggregate the generalized area class soil survey information into spatially explicit raster domain data. One of the six steps of the DSMART procedure is to use legacy soil survey-derived pseudo points to sample underlying covariate information. The sampled information in turn is used for prediction and uncertainty evaluation (Odgers et al., 2014; Geng et al., 2016).

2.5 Applying topographic and remotely-sensed data in PSM

Although there are many predictors, covariates or features that can be used for PSM, not all of them are equally important and effective in practice (Nussbaum et al., 2018). However, topography plays an important role in the development of soil and its classification, especially in Canada where the classification system and soil surveys have been based on the theory of soil genesis (Jenny, 1941). For predictive statistics-based PSM, topographic related attributes such as local scale morphometry (e.g. slope gradient, curvature and aspect etc.), landscape scale morphometry (e.g. multi-resolution ridge top flatness and valley bottom flatness indices) (Gallant and Dowling, 2003), and hydrological properties (e.g. topographic wetness index and slope length factor etc.) are often derived from digital elevation models (DEMs). In Canada, there are several accessible DEM sources such as the SRTM (Shuttle Radar Topographic Mission) 30 m grid DEM, ASTER- (Advanced Space-borne Thermal Emission and Reflection Radiometer)-based 30m grid DEM, legacy Canadian Digital Elevation Model (CDEM), and sporadic finer (<10 m) grid DEMs from airborne sensors. It is anticipated that finer resolution (less than 5 m grid size)

DEMs of Canada will become more publicly accessible in the early 2020's. DEMs along with the derived landscape facets are especially necessary for PSM in mountainous and undulating landscapes in Canada. In level or plain areas such as the Red River basin (Liu et al., 2012), finer resolution DEMs and their derivatives are necessary, such as those acquired using LIDAR (Light Detection and Ranging).

Across the diverse Canadian landscapes, there have been some PSM efforts in mountainous regions (MacMillan et al., 2000, 2007; Smith et al., 2010; Bulmer et al., 2016; Heung et al., 2016), in undulating areas (Niang et al., 2008; Zhao et al., 2008a; Yang et al., 2011; Nyiraneza et al., 2017), in low-relief agricultural regions (Liu et al., 2008; Liu et al., 2012), in forested regions (Beguin et al., 2017) and at a national scale (Hengl et la., 2017). MacMillan (2007) produced an ecological-landform map on 3 million ha of forested land in British Columbia, Canada using fuzzy and boolean logic with various environmental variables and expert knowledge. MacMillan et al. (2000) developed a computer program called LandMapper, a heuristic, fuzzy-logic based landform model which is an extension of a seven-unit hillslope model described by Pennock et al. (1987, 1994). Interestingly MacMillan et al. (2000) concluded that it is not the absolute accuracy of the DEM that most influences the results of landform classification, instead it is the resolution of a smoothed abstraction of the landscape at the scale of interest to the classifier. They found that it is necessary to filter most DEM data several times to reduce the effects of local noise. Zhao et al. (2008a, 2008b) applied a neural network inference approach for soil texture and soil drainage class mapping in an undulating maritime region in Canada. Again, topographic derivatives were shown to be an important part of the inputs. Nyiraneza et al., (2017)

published time series soil organic carbon (SOC) maps using gridded sampling data and top ranked environmental covariate to digitally map SOC changes over an 18-year time span. Hengl et al. (2017) used Random Forest machine learning along with globally collected legacy soil pedon data and environmental covariates, to produce global 250 m soil grids. The validation of the 250 m grid of the Canadian portion is currently being conducted (He et al., 2017).

For areas where landscapes are generally level or slightly undulating, the challenge of PSM is that easy-to-observe factors such as landform and vegetation conditions cannot effectively reflect soil spatial variations (Mendonca Santos et al., 2000; Odeh and McBratney, 2000; Iqbal et al., 2005). It becomes even more difficult when the needed surficial geological material data are also lacking or too coarse. One of the possibilities to overcome this problem is to use multi-spectral and time series (optical) remotely-sensed data (Hengl et al., 2004; Zhu et al., 2009; Liu et al., 2012). This approach is based on the hypothesis that the differences of sensed spectral signatures before and after major rainfall events are partially related to soil properties such as texture, which vary in space. However, it can be difficult to acquire good quality optical remotely-sensed data right after major precipitation events due to cloud as well as vegetation cover. Fortunately, short vegetation and cloud cover is not an issue for Synthetic Aperture Radar (SAR) sensors. In plains or generally level agricultural and grassland areas, SAR-derived backscatter coefficients have been used for soil drainage class mapping (Niang et al., 2008; Liu et al., 2008). Recently,

Geng et al. (2010a) discovered that decomposed polarimetric RADARSAT-2 fine beam backscatter data are strongly correlated with soil surface texture.

Although the performances of some PSM algorithms have been studied in Canada (Heung et al., 2017), the relationship between the outputs of frequently used algorithms and the scales or resolutions of the features or covariates used in PSM still need to be properly studied for the Canadian context. Behrens et al. (2010, 2019) concluded that soil distribution can be best predicted by a combination of features including DEM derivatives at different scales and only a small number of predictor features are typically required to achieve a good prediction. To rank and choose effective features and resolutions, simple linear (e.g. ANOVA based F test) approaches can be used. Multiple feature ranking and importance evaluation are also realized using the Best General Linear Model (BGLM) (Xu and McLeod, 2009) and ensemble machine learners such as RF. Lindsay (2016) has developed algorithms and procedures to evaluate and select those significant land morphometry derivatives to avoid broadly using large mixes of numbers and scales of geospatial features for modeling.

2.6 Covariate features selection, feature scales and PSM accuracy

The use of environmental covariates in the context of a soil genesis model is common in PSM. However the accuracies of these predictions are often not as high as expected due to the lack of quality training information for machine learners (Minasny and McBratney, 2016). In addition, the scales and efficiency of the features used for the soil type or soil property inference can also greatly impact the outcomes of PSM. For example, using high

resolution DEM-derived slope gradient data to map soil types may only be effective at certain scales or resolutions. During conventional soil surveys, soil surveyors would estimate slope gradient by referencing area slope length and overall gradient to delineate soil type boundaries, so the DEM-derived gradient information had to match the scale of the generalized soil gradient knowledge. Similarly, when surficial geological material data are used as a surrogate for soil parent material information, the availability and quality of an independently produced surficial geological material map can greatly influence the outcomes of PSM. It is important to analyze and understand those uncertainties and find alternative solutions for PSM. In the situation when topographic and surficial geological material data are not suitable for the desired precision of PSM, other covariates, especially features derived from remotely-sensed data may be feasible.

Chapter 3: **Selected inference algorithms for predictive soil mapping**

3.1 Introduction

After reviewing the state and challenges of current Canadian soil survey programs and the national soil data holdings, this chapter summarizes basic concepts, theories and uses of several frequently-used inference methods such as classification tree, neural networks and knowledge-based fuzzy logic sets, in the context of soil class and soil property prediction. These machine learning methods are just some of the many approaches used in the field of PSM in the last two decades. Although it is not the focus to reiterate the theories and mathematics of the inference algorithms, the reviewed inference methods should not be treated as "black boxes" either. For additional illustration, practical examples of applying the algorithms are provided in Appendix A.

3.2 Inference and knowledge based predictive soil mapping

The purpose of classification and inference analytics is to mathematically predict or infer soil class and/or soil property distributions. There have been many algorithms for predicting continuous or categorical variables from a set of continuous or categorical factors or predictors (Cianfrani et al., 2018). For example, using GLM (General Linear Models) and GRM (General Regression Models), we can construct relationships between continuous and categorical factors to predict a continuous dependent variable; using GDFA (General Discriminant Function Analysis), we can solve the classification problem of categorical variables. Statistical models such as GLM, generalized additive linear model (GAM), ordinary kriging (OK), and regression kriging (RK) have been frequently used in

PSM-related works (McBratney et al., 2003; Hengl et al., 2015; Beguin et al., 2017). As for inference methods, machine learners such as random forest (RF) (Breiman, 2001), Cubist (C5) (Quilan et al., 1993), artificial neural networks (ANN) (LeCun, 2015), support vector machine (SVM) (Mondal et al., 2012), and Bayesian models (Beguin et al., 2017) have been widely tested and used in PSM.

3.3 Machine learning using tree methods

Decision tree frameworks and their principles are fundamental to these above-mentioned data mining programs. A decision tree is an inverted tree composed of a root node from the top, a set of interior nodes, and terminal nodes called leaf nodes at the bottom. A node is linked to a classification process which is implemented by a set of rules that determine the structure of a classification tree. The leaf or terminal nodes represent the final classification (Tso and Mather, 2009). Within classification and regression tree (CART) analysis, the process for predicting continuous dependent variables is called regression, and for categorical predictor variables is called classification. The key difference between CART and other tree algorithms such as C5 (Quinlan, 1993) is that CART uses a binary classification scheme. It can only allow two branches or two nodes to form at each splitting stage (Tso and Mather, 2009). Recent studies have compared strengths and weaknesses of tree-based classifiers as well as other inference algorithms (Heung et al., 2017; Hengl et

al., 2017). Among the tree-based classifiers, the RF algorithm has gained popularity in the field of predictive soil mapping.

In order to "grow" a decision tree effectively and accurately, recursive evaluation of classification performance with classification rules is conducted using evaluation indices, namely information gain (Quinlan, 1993) and the Gini impurity index (Breiman et al., 1984). These are used in tree algorithms to determine if a node should be split.

3.3.1 Information gain

During classification tree induction process, to split or not to split a node on an attribute or feature (e.g. slope gradient) can be evaluated with the information gain index ($I_E(t)$) as detailed in Equations 3.1, 3.2 and 3.3 below. The attribute or feature with the greatest information gain is chosen for splitting a node. To further explain how the information gain is calculated and used, Appendix A.1 presents modified working examples based on Tso and Mather (2009).

$$I_E(t) = -\sum_{j=1}^{m} f(t,j) \log_2 f(t,j) \qquad \text{Equation 3.1}$$

where $f(t,j)$ is the proportion of training samples belonging to class j, $j \in \{1,2,\ldots,m\}$, within node t, and m is the number of classes. If node t contains N_t samples, then $f(t,j)$ is calculated by the following expression:

$$f(t,j) = \frac{1}{N}\sum_{i=1}^{N_t}\Gamma(y_i, j) \qquad \text{Equation 3.2}$$

where

$$\Gamma(y_i, j) = \begin{cases} 1, if \ldots . y_i = j \\ 0, otherwise \end{cases} \qquad \text{Equation 3.3}$$

3.3.2 Gini impurity index

The Gini impurity index measures the impurity of an input feature (e.g. land use type) with respect to the classes. Gini impurity index approaches its minimum when all attributes in a node fall into a single class, and is typically calculated using Equation 3.4. Again, more details about this index and its application can be found in Appendix A.1.

$$I_G(t_{X(x_i)}) = 1 - \sum f(t_{X(x_i)}, j)^2 \qquad \text{Equation 3.4}$$

where

$f(t_{X(x_i)}, j)$ is the proportion of samples with the value x_i belongs to class j at node t. The decision tree splitting criterion is based on choosing the attribute with the lowest Gini impurity index.

3.3.3 Random Forest tree-based machine learning

RF, unlike ANN, SVM, or Bayesian networks (BN), is a supervised ensemble classifier that divides a complex classification process into a series of decision processes (i.e. trees).

With the RF algorithms, multiple decision trees are trained and the results are based on the predictions from an ensemble of the individual trees (Breiman, 2001). Each individual tree is generated by applying a subset of covariates (i.e. features) that is randomly sampled from the total number of covariates on a bootstrapped sample of training data. Bootstrap sampling is conventionally done by using 70% of the total samples for classification and 30% for internal validation. The most frequently used feature selection measures are the Information Gain (Quinlan, 1993) and the Gini Index (Breiman et al. 1984). In RF algorithms each tree or classifier is grown to the maximum extent with no pruning so that an ensemble of trees is created. In the context of raster data-based classification, each pixel in the final classification map is assigned to the majority class across all the tree predictions of the pixel in the forest. The majority rule is done by employing a voting procedure of selecting the most popular class for the pixel. This ensemble approach also produces associated probability or uncertainty ratings for each classified pixel. RF uses recursive partitioning to evaluate the relationship between the participating co-variables and variable of interest (Breiman et al., 1984). RF can handle multiple data types, missing data, nonlinear relationships, outlier data and dependent variable interactions effectively. The RF classifier is gaining increasing attention because of its fast computation, few required parameters, few statistical assumptions on training data, low sensitivity to noise or over-fitting, and the capability to determine variable importance (Pal, 2005; Rodriguez-Galiano et al., 2012; Harris et al., 2014; Zhang et al., 2017).

3.3.4 Use cases of tree-based predictive soil mapping

The quality of tree-based algorithms to predict soil classes and properties has been limited by several key issues such as source and quality of training data and the kinds and scales of environmental covariates. The uses discussed in this section have addressed these key issues either individually or collectively. Moran and Bui (2002) used a data mining tool called C5 (Quinlan, 1993) as the rule induction engine to build decision trees while studying strategies to improve PSM quality. Existing soil survey data were used to construct decision tree models and rules. During the induction operation, techniques of tree pruning and boosting were implemented. The purpose of tree pruning is to simplify the tree models. The boosting measure improved quality of input information capture. Among the predictive attributes used in the study, terrain attributes were marginally more powerful than the Landsat MSS data. Furthering the work of Moran and Bui (2002), Schmidt et al. (2008) systematically analyzed the effect of instance or sample size reduction on decision tree classification accuracy. With 95,000 pedons or instances, they tested the impact of proportional and disproportional stratified random sampling on the decision tree outcomes. The study showed that appropriate sampling in combination with a grid search method predicted soil more accurately than those based on a non-instance selection grid learning method.

Henderson et al. (2005) used the national soil point database and selected predictors and environmental variables to construct decision trees to predict soil pH, organic carbon, total phosphorus, total nitrogen, thickness, particle (clay, silt and sand) contents Australia-wide. The study found that environmental variables even at 250 m grid resolution can be used as

predictors for soil property mapping in their case. However substantial point data are required in order to construct the decision trees effectively. The decision tree program that they used, Cubist, available from http://www.rulequest.com/ (Accessed May 7, 2020), uses approaches similar to regression trees in CART (Breiman et al., 1984) except that it is based on linear models.

Scull et al. (2005) studied how the variations of dependent variable (soil classes) grouping, soil predictor (environmental) variables and mapping area segmentation influence the outcomes of PSM. Through the study, they concluded that grouping soil classes at a soil great group level is recommended for the predictive soil mapping in the study area. Landform variables explained the distribution of soil types directly. Dividing the study area into sub-units increased the accuracy of the decision tree-based prediction.

Kheir et al. (2010) used systematic classification tree models to predict the spatial extent and distribution of organic carbon in soils at national scale in Denmark. In the study, seventeen parameters (e.g. parent material, soil type, landscape type, elevation, slope gradient, slope aspect, mean curvature, plan curvature, profile curvature, flow accumulation, specific catchment area, wetness index, NDVI) were all or partially used to construct decision tree-models. The constructed tree models all produced above 68% accuracy on soil organic carbon (SOC) at 1:50,000 scale.

Recently Hueng et al. (2014, 2017) conducted soil parent material and soil class mapping using both legacy soil survey data and multiple machine learning algorithms in western

Canada. Among the machine learning algorithms, the ensemble approaches with bagging methods produced above 60% prediction accuracy. The combination of bagging and ensemble machine inference also produced uncertainty information for the digital soil maps. Hengl et al. (2017) mainly focused on the utility of RF-based inference, producing a global soil class and soil property map at a resolution of 250m. Although the 10-fold internal cross validation was satisfactory, regional independent validation demonstrated low accuracy of predicted soil organic carbon in Canada (He et al., 2017). That was partially due to the limited amount of soil point data that were used during the global calculation. In locations where more soil point data are available, soil properties (e.g. soil organic carbon and soil total phosphorus) can be mapped with RF directly with high accuracy (Nyiraneza et al., 2017; Benjannet et al., 2018).

3.4 Artificial Neural networks and their application in predictive soil mapping

3.4.1 Neural network structure

To solve prediction and analytic problems, ANNs are widely used in remote sensing, GIS and PSM. Pattern recognition is perhaps the most common use for ANN. Recent advances and applications of Convolution Neural Networks (CNN) in the field of machine learning open new frontiers of predictive soil mapping using deep learning approaches (LeCun, 2015).

An artificial neural network consists of interconnected cells called neurons. Neurons can process incoming information and activate other connected neurons to continue the process. Similar to a human's brain, the individual neurons that make up an artificial neural

network are interconnected. These connections allow the neurons to signal each other as information is processed. In the case of ANN, not all connections are equal; each is assigned a connection weight. If there is no connection between two neurons, then their connection weight is zero. Groups of networks come together to form layers. The connection weights determine the output of a neural network layer and ultimately of the neural network.

3.4.2 The neuron

A neuron is a communication conduit that both accepts input and produces output. When a neuron produces output, the neuron is said to activate, or fire. A neuron will activate when the sum of its inputs satisfies the neuron's activation function. Consider a neuron that is connected to a number of other neurons. The variable "w" represents the weights between this neuron and the other "*k*" neurons. We will say that this neuron is connected to "*k*" other neurons. The variable "*x*" represents the input to this neuron from each of the other neurons. Therefore, we must calculate the sum of every input "*x*" multiplied by the corresponding weight "*w*" as indicated below:

$$u = \sum_{k} w_k x_k \qquad \text{Equation 3.5}$$

This sum must be used in a threshold function to evaluate whether the neuron should fire or not. There are several threshold methods that are commonly used by neural networks. The simplest one is to use a specific range of values (e.g. threshold range between 5 and 10); when dealing with a range of numbers both greater and less than zero, a hyperbolic tangent threshold should be used. If only a positive data range is needed, then the Sigmoid

threshold method is used.

$$y(\mu) = \frac{1}{1 + e^{-\mu}} \qquad \text{Equation 3.6}$$

A modified Sigmoid function can handle both positive and negative data situations. This is realized by using TANH function as detailed in following equation:

$$y(\mu) = \frac{e^{\mu} - e^{-\mu}}{e^{\mu} + e^{-\mu}} \qquad \text{Equation 3.7}$$

3.4.3 ANN layers and training

Heaton (2005) provide a classic text on this process and the explanation in this section is based on that source. The neuron connection weights are what give a neural network the ability to recognize certain patterns. By adjusting the weights, the neural network will recognize a different pattern. In fact, the process of training is to adjust the individual weights between each of the individual neurons until the desired output is achieved.

Neurons are usually grouped into layers, which are groups of neurons that perform similar functions. There are three types of layers: the input layer is the layer of neurons that receive input from the user/program; the layer of neurons that send data to the user/program is the output layer; between the input and output layer can be zero or more hidden layers. Hidden layer neurons are connected only to other neurons and never directly interact with user/program (Figure 3-1).

In order to process information and predict outcomes, an ANN needs to be trained either through unsupervised or supervised training. Most training algorithms begin by assigning random numbers to the weight matrix. Then the validity of the neural network is examined. Next, the weights are adjusted based on how valid the neural network performed. This process is repeated until the validation error is within an acceptable limit.

For supervised training, users need to provide sampled inputs and anticipated output (e.g. environmental covariates and soil classes). The expected outputs will be compared against the actual output from the neural network, and the 'back-propagation' training process will include calculating errors between the network outputs and the anticipated outputs, and adjusting the weights of various layers backwards from the output layer all the way back to the input layer Heaton (2005).

Unsupervised training usually occurs when the neural network is used to classify the inputs into several groups. A common application for unsupervised training is data mining. In this case, one has a large amount of data, but does not often know exactly what one is looking for. ANN is used to classify the data into several groups. Unsupervised training is a very common training technique for Kohonen neural networks, as discussed below Heaton (2005). A trained neural network has to be validated before actual use. To correctly validate a neural network, validation data must be set aside that is completely separate from the training data.

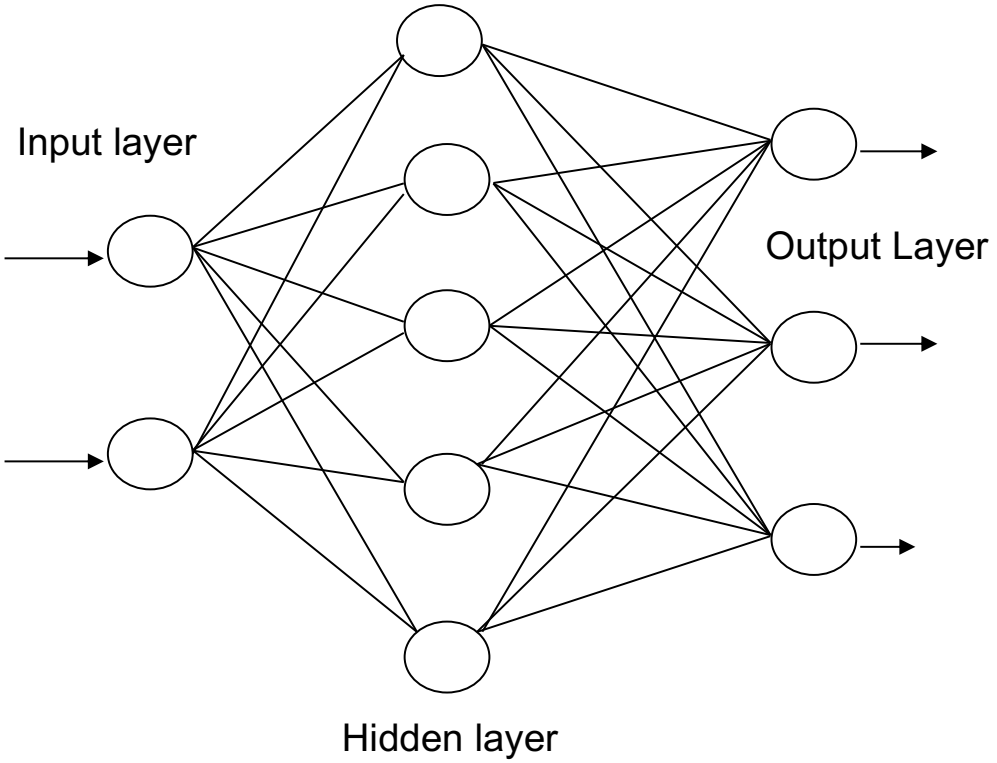


Figure 3-1 A typical three layer artificial neural network (ANN) (based on Heaton, 2005)

3.4.4 Types of ANN

Not every neural network has to have all the three layers as presented in Figure 3-1; the input and output layers are required, but it is possible to have one layer act as both as an input and output layer, with no hidden layer. The Hopfield neural network is a single layer neural network (Figure 3-2). Kohonen Neural Networks, sometime called self-organizing feature maps (SOFM), are ANNs with no hidden layer. A Kohonen neural network neither uses any sort of activation function nor has bias weight. It is trained with an unsupervised training method. Since it only has two layers, input and output, it can only handle linear issues. The advantage of a Kohonen Neural Network is that it is relatively simple to construct and can be trained very quickly. Please refer to Appendix A.2 for a working

example to examine how a Kohonen network is structured, trained and used (see also Heaton, 2005, page 157-162).

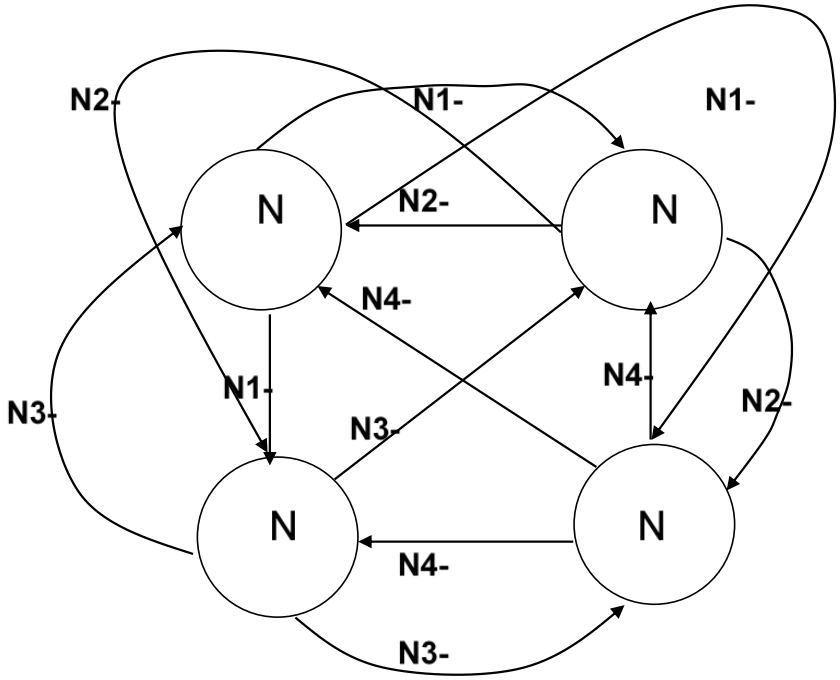


Figure 3-2 Hopfield neural network with 12 connections (based on Heaton, 2005)

A feed-forward back-propagation neural network is one of the most common neural network architectures. Each layer of a feed-forward back-propagation neural network contains connections to the next layer, but there are no connections back. "Back-propagation" indicates that this type of neural network requires supervised training. The "back-propagation" and "feed-forward" algorithms are often used together. The structure of a typical feed-forward neural network often includes an input layer, one or more hidden layers and output layers. Although there can be more than one hidden layer in a feed-forward neural network, it is rare to have more than two hidden layers.

Deciding how many neurons are needed in a layer is important for a feed-forward neural network. Too many neurons or too few neurons can cause a feed-forward neural network to either over fit or under fit. The rule-of-thumb is: the number of neurons in the hidden layer should be 2/3 of the number of neurons of the input layer plus the number of neurons from the output layer; the number of the neurons in the hidden layer should be less than twice the input layer size.

3.4.5 ANN and predictive soil mapping

Using ANN for PSM also faces training data and environment covariate selection issues. The advantage of ANN is its capacity to handle nonlinear soil-environment relationships. However, different kinds of learning and inference procedures even within the ANN domain can produce contrasting outcomes as detailed through cases in this section. The back-propagation based ANN or deep learning algorithms shows promise in PSM as well (Zhao et al., 2009; Poggio et al., 2019).

To meet hydro-ecological modeling needs, Zhu (2000) used multilayer feed-forward neural networks to map soil classes. The input of the network was a set of selected environmental covariates; the output of the network was a set of similarity values linked to prescribed soil classes. The study showed that the soil class map derived using the neural network method is much more detailed and of higher quality than that produced from the conventional soil survey.

Chang and Islam (2000) explored the possibility of inferring soil physical properties using ANN with multi-temporal remotely-sensed passive microwave brightness temperature data and soil moisture maps as inputs. The research was based on the assumption that the dry-down curves of brightness temperature and soil moisture at different locations with similar soil texture would exhibit similar trends. The study concluded that the Feed-Forward Neural Network (FFNN) with supervised training provided better outcomes (soil texture maps) than that based on the SOFM. Chang and Islam (2000) reported that using remotely-sensed soil moisture data as input along with known soil texture information of some locations, the FFNN approach provided the best outcome, and more accurate soil texture maps.

The utility of the ANN allocation procedure has also been studied using the Czech soil survey of agricultural land data (Boruvka and Penizek, 2006). In the study, terrain attributes such as elevation, slope aspect, and slope were used as covariates to train a Perceptron type of neural network. To improve the classification accuracy, a bootstrap aggregating (bagging) method was introduced into the neural network training process. Boruvka and Penizek (2007) stressed that number of input neurons does not have to be large. Selecting distinguishable classes and input neurons can improve the classification or allocation outcomes. Good training data and suitable neural network structures are important. However, the fact that the Perceptron neural network is incapable of solving non-linear problems (Heaton, 2005) was not addressed.

In Canada, Zhao et al. (2009) constructed a back-propagation artificial neural network along with input variables such as soil terrain factor, sediment delivery ratio, and vertical slope position, to predict soil texture distribution. The relative accuracies for clay and sand content were 88% and 81% respectively. The study suggests that the constructed ANN can be extended to other areas where the range of input parameters is similar to the area where the model was calibrated. Zhao et al (2008a) extended their ANN model for soil organic carbon content mapping in the same area. The predicted soil organic carbon values had 87% of the total points within the ±0.5% of the measured values and 98% of the total points within ±1% of the measured values.

3.5 Support Vector Machine

SVM, another frequently used machine learning methods in PSM, uses a subset of training data called support vectors to partition the feature space of co-variables without requiring specific statistical distributions of the training data. SVMs use optimization algorithms to generate partition boundaries or hyperplanes that separate classes within a selected feature space, or between co-variables. Contrasting with the low user input requirements of RF algorithms, SVMs require the user to specify a set of parameters such as the choice of kernel, kernel specific parameters, and regulation parameters. The choices of those parameters often have great influence on the outcomes of the inference (Huang et al., 2002; Pal, 2005; Mountrakis et al., 2011; Mondal et al., 2012), so optimum parameters need to be selected in order to achieve the best classification result (Kavzoglu and Colkesen, 2009).

3.6 Fuzzy sets and their application in predictive soil mapping

3.6.1 Fuzzy sets overview

The objective of classification is often to allocate features such as soil types into classes or groups. Classes with crisp or clear boundaries are called crisp sets; classes with overlapping boundaries are called fuzzy sets. A feature or element of a crisp set can only be a member of one mutually exclusive group (membership grade 1 or 0). By contrast, a feature or element of a fuzzy set can have partial membership grades in any group. Both crisp and fuzzy sets have advantages in certain circumstances.

Let $\mathbf{X} = \{x\}$ represent a finite set or space of points such as elements, objects or properties. A fuzzy subset **A** of **X** is determined by a fuzzy membership function (FMF) μ_A. For each x, the FMF $\mu_A(x)$, defines the membership grades (within 0 and 1) of fuzzy objects in the ordered pairs:

$$\mathbf{A} = \{ x, \mu_A(x)\} \text{ for each } x \in \mathbf{X} \qquad \text{Equation 3.8}$$

The membership grade is expressed by

$$\mu_A(x) \rightarrow [0,1] \qquad \text{Equation 3.9}$$

The crisp set can be seen as a special case of the FMF. When the membership grade equals 0, we say x does not belong to the subset **A**; when the membership grade equals to 1, x fully belongs to the subset **A**; when $0<\mu_A(x)<1$, partial membership is defined.

A fuzzy member is defined as normal or convex. When the maximum value of membership

in a fuzzy set is 1, the fuzzy numbers are said to be normal; when the membership follows patterns of increasing, decreasing and/or leveling off, the fuzzy numbers are convex. Mathematically, the convex fuzzy membership can be expressed for each of the real number *a*, *b* and *c* as:

$$\mu_A(b) = \min(\mu_A(a), \mu_A(c)), \quad a<b<c \qquad \text{Equation 3.10}$$

Theoretically, FMFs can take any form as long as the function can provide a membership grade within 0 and 1. There are four main types of frequently used FMFs: monotonic, triangular, trapezoidal, and bell-shaped (Figure 3-3).

For a three element set as expressed as $\mathbf{X} = \{0, 5, 27\}$, assume the FMF for fuzzy subset A is defined by

$$\mu_A(x) = \frac{10}{0.5x^2 + 10} \quad x \in \mathbf{X} \qquad \text{Equation 3.11}$$

The fuzzy subset **A** can be represented as

$$\mathbf{A} = (1/0) + (0.44/5) + (0.027/27)$$

Note: here 0.027/27 means that the element with value 27 has a membership grade of 0.027 in **A**.

The height of a fuzzy subset **A**, height(**A**), is defined as the highest membership value contained in that fuzzy subset. In above example, the height of the fuzzy subset **A** is height (**A**) = 1.

3.6.2 Fuzzy C-mean classification

Fuzzy set theory provides solutions to dealing with uncertain and, especially, imprecise boundaries between mapped soil categories (Burrough et al, 1997; McBratney and Odeh, 1997). Membership in soil classes is generally derived in two ways: fuzzy c-means (FCM) and the Semantic Import model (SI) (Burrough et al., 1992a, 1992b; Zhu et al., 2010). FCM is a data-driven unsupervised approach; the SI-based approach uses fuzzy membership functions which are selected along with a priori expert knowledge. In SI-based approaches, the process of constructing expert knowledge and the quality of acquired knowledge can all impact the outcomes of fuzzy logic inference, as will be further discussed later.

Under fuzzy logic, assume that there are p point values of n soil variables. This forms a data matrix, **X**, of size n x p values which may be grouped into c classes and therefore produces n x c matrix, **U**, of membership values. Given $\mathbf{U} = u_{ik}$, $u_{ik}=1$ if a site or individual has full membership to class j. The following conditions are exclusive and jointly exhaustive:

$$\sum_{i=1}^{c} u_{ij} = 1 \qquad 1 \leq k \leq n \qquad \text{Equation 3.12}$$

$$\sum_{k=1}^{n} u_{ik} > 0 \qquad 1 \leq i \leq c \qquad \text{Equation 3.13}$$

$$uk \in \{0,1\} \qquad 1 \leq k \leq n;\ 1 \leq i \leq c \qquad \text{Equation 3.14}$$

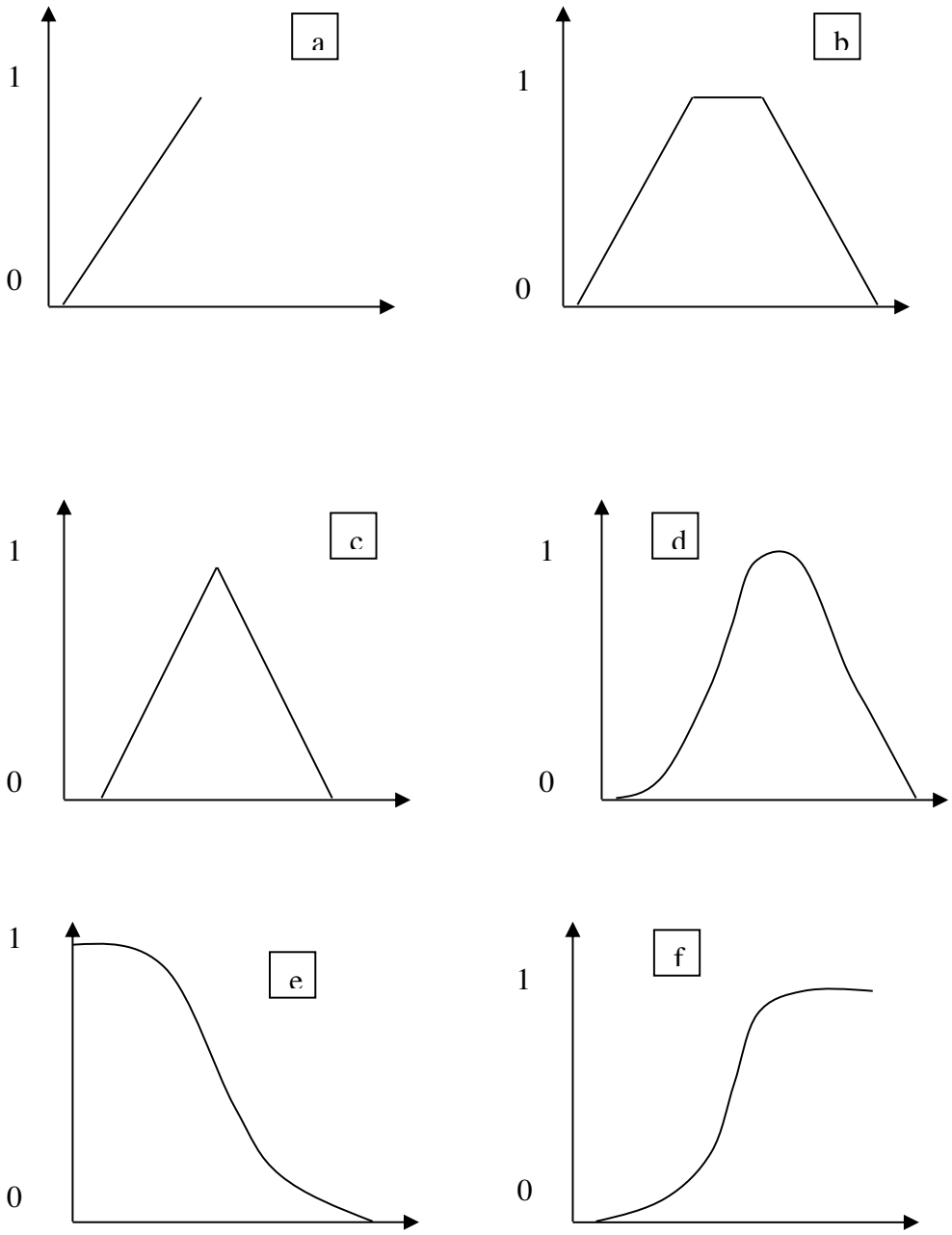


Figure 3-3 Frequently used fuzzy membership functions. a–monotonic; b- triangular; c- trapezoidal; d-bell-shaped; e- reverse S shaped; f-S shaped

Similar to hard-c means classification, the FCM aims to minimize the within-class sum-of-square errors (J_m) in the objective function below (Equation 3.15) (Odeh et al., 1992a, 1992b) while satisfying the above conditions of Equations 3.12, 3.13, and 3.14.

$$J_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d_{ik}^2$$

Equation 3.15

$$d_{ik}^2 = \|y_k - v_i\|_A^2 = (y_k - v_i)^\tau A(y_k - v_i).....y \in Y$$

where ***U*** is the fuzzy membership value matrix; v is the centroidal value of a feature class; ***Y*** is the variable value matrix; n is the number of variables or features; c is the number of classes or groups; m is fuzziness exponent ($1<m<\infty$); d_{ik} is the weighted distance between y_k and centroid value of v_i, u_{ik} is the membership values of an attribute value at k^{th} point for class i; **A** is weight matrix; and J_m is fuzzy classification error.

Methods for choosing the number of classes, "c", and the value of the fuzziness exponent (m) are important when conducting FCM classification. One technique is to use the partition coefficient, **F**, and normalized entropy, **H**, to optimize the number of classes, "c", as expressed below.

$$F_c\left(\hat{u}\right) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(\hat{u}_{ik}\right)^2 / n$$

Equation 3.16

$$H_c\left(\hat{u}\right) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(\hat{u}_{ik}\log\left(\hat{u}_{ik}\right)\right) / n \qquad \text{Equation 3.17}$$

where *F* measures degree of overlap among classes; *H* measures fuzziness of classes. When the number of classes, "*c*", increases, *H* increases and *F* decreases. When increasing or decreasing class numbers produce minimum changes of *H*, an optimum number of classes, "*c*", can be reached. Certainly this optimum number changes when the fuzziness exponent, m, changes. It is recommended that the m value should be between 1.5 and 2.5 (Odeh et al., 1992a, 1992b).

3.6.3 Soil inference under fuzzy logic similarity framework

Using fuzzy logic concepts, soil at a location (or pixel) can be assigned to more than one soil class with varying degrees of class membership; it can bear a partial membership in each of the prescribed soil classes. Each fuzzy membership is seen as a similarity measure between the local soil and the typical or central case of the given class. All fuzzy memberships are retained in a similarity vector with n elements (Figure 3-4), where n stands for the number of identified soil classes and S_{ij}^{k} in the vector represents the similarity value between the soil at pixel (i,j) and the central case of soil class "k".

With a raster data model, soil of an area can be represented as an array of pixels with soil at each pixel being represented as soil similarity vector (Figure 3.4). In this manner, soil spatial variation is represented as a continuum in both the spatial and parameter domains (Zhu et al., 1997). Appendix A.3 presents an explicit example of fuzzy membership arrangement for a hypothetical soil catena.

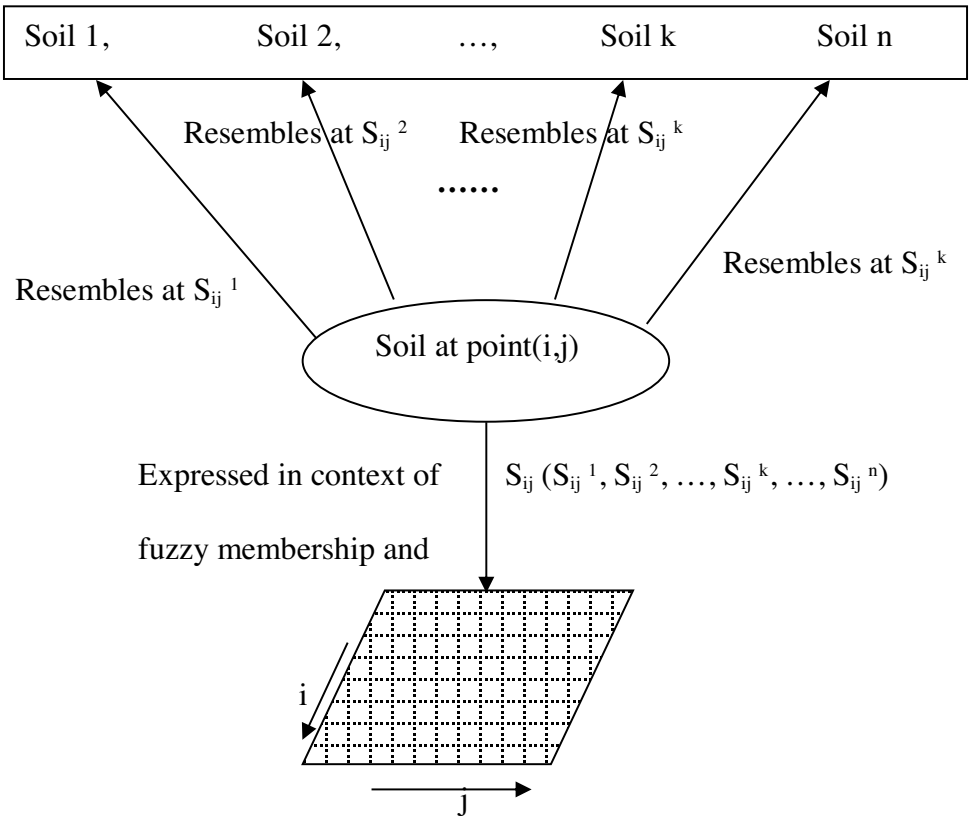


Figure 3-4 Illustration of fuzzy membership of soils in raster domain (based on Zhu et al., 1996)

3.6.4 Predictive soil mapping with fuzzy logic

Fuzzy logic framework FCM algorithms have been used to map soil classes with or without training data sets (Bezdek et al., 1984). Earlier applications of FCM on soil classification were reported by Odeh et al. (1992a, 1992b), applying FCM and producing fuzzy membership values to guide purposive soil sampling. The purposively sampled point data were further used with kriging for a hardened (crisp) soil class map. In those studies, crisp maps were derived by hardening the fuzzy membership values.

In the SI model domain, the relationship between soil and soil forming variables are constructed using 'central concepts' of soil classes defined through the soil classification system (McBratney and Odeh, 1997; Qi et al., 2006). The relationship can be expressed with a priori membership FMFs. One frequently used FMF is depicted by the symmetrical bell-shaped curve (Figure 3-3 d) (Burrough et al., 1992).

Zhu et al. (1996) furthered the SI-based fuzzy logic classification by feeding expert knowledge acquired from existing soil survey and local soil experts into an inference engine called soil-land inference model (SOLIM) with a user-friendly computer graphic user interface (GUI). In the SOLIM's similarity scheme, the soil type at a given pixel is a vector of similarity values. With expert knowledge and selected environmental variables as inputs, the method can infer soil series from environmental conditions accurately (Shi et al., 2004; Qi et al., 2006; Zhu et al., 2010). Shi et al. (2004) studied the effectiveness of case- or location-based knowledge reasoning for digitally mapping soil classes and properties. With an improved SOLIM program, case-based knowledge was collected through interactions between soil experts and a GIS-based GUI. The key advantage of case-based reasoning is that it does not need to assume the independency of environmental variables. The derived raster soil map using case-based reasoning had higher accuracy than that from the conventional survey (Shi et al., 2004). Qi et al. (2006) used the prototype theory from the expert system domain to define and acquire soil expert knowledge of prescribed soils or prototypes. The study summarized the advantages and disadvantages of prototype (rule)-based and case-based reasoning approaches. In order to find a middle

ground with fuzzy logic, Shi et al. (2009) developed a program called a soil inference engine (usually referred to as SIE) to handle the integration of the two reasoning measures. Although the combined reasoning approach produced satisfactory outcomes, further tests and studies are suggested.

The real challenge to use fuzzy logic knowledge to map soils is how to quantify the descriptive knowledge in the form of membership functions (Zhang et al., 2017). A further challenge also arises from mapping projects containing many members or soil types. The overlapping and fuzziness among the soil types results in higher uncertainties of output soil maps. Zhu et al (2010b) studied the situation in which soil expert knowledge and conventional soil survey information are both absent. The study used purposive soil sampling (Zhu et al., 2008) with FCM to construct expert knowledge and then to infer the soil class and soil A horizon carbon content under the fuzzy logic system. Purposive soil sampling aims to sample soil points cost-effectively while maintaining statistical requirements for representation. This was realized by including soil genesis-related environment co-variables. The predicted soil class and soil carbon maps were more accurate than the ones produced with a linear regression model.

In conclusion, FCM is useful for spatial pattern analysis and purposive soil design in the context of PSM, but fuzzy logic-based inference using fuzzy membership functions can only be effectively used for PSM with limited soil types. Too many soil types will have too many overlapping fuzzy membership functions that reduce the inference accuracy.

Chapter 4: **Predictive mapping of soil classes using legacy soil surveys**

4.1 Introduction

In the last twenty years, PSM has undergone rapid development to the point where it now has become operational through cost effective production of raster soil data maps (McBratney et al., 2003; Heung et al., 2016; Arrouays et al., 2017; Hengl et al., 2017; Zhang et al., 2017). Point soil data are essential for both numerical modeling-based and knowledge-based PSM (Adhikari et al., 2014). When recent field point soil data are lacking or insufficient, legacy soil point data from soil profile databases or those derived by data mining procedures become vital. However, legacy point data have several limitations: many of the legacy point datasets are not publicly accessible; legacy soil point data are not sampled under strict statistical requirements and therefore are often spatially skewed or clustered; methods of soil property measurements are not consistent; and data are collected over large ranges of time spans. Regardless, legacy soil survey data mining has been utilized extensively in PSM (Dobos et al., 2010; Vaysse and Lagacherie, 2015; Bulmer et al., 2016; Heung et al., 2014, 2017; Liu et al., 2016, Vincent et al. 2018, Zerratpisheh et al., 2019). In Canada, where nationwide soil point data are sparse (Geng et al., 2010b), detailed soil surveys from 1:10,000 to 1:65,000 exist in some agricultural regions. Detailed soil surveys can be used to extract point soil information and related knowledge for PSM, and are often at the level of detail such that many mapped soil polygons contain only one soil component. A soil component in the national soil database in Canada means a soil type mapped and described in a map unit or soil polygon. A mapped soil polygon can have more than one soil component. A soil map polygon with a single component indicates that only

one major soil exists in that mapped area. Extracting knowledge from even a single component soil polygon of legacy soil surveys requires the assumption of a homogenous soil polygon. As a consequence of this, I hypothesize that any location within a single-component soil polygon can be used to represent the reported soil component or type. The primary goal of this study was to test this hypothesis by assessing the accuracy of a number of data mining and machine learning methods using the detailed 1:20,000 soil surveys of Prince Edward Island, Canada. With the availability of the 1:20,000 soil surveys and the permission to sample and inspect soils in the field, a site in PEI was selected.

4.2 Site description

This study is conducted in the province of Prince Edward Island (PEI), Canada, in the Gulf of St. Lawrence. The island is crescent-shaped, about 230 km long, 6 to 30 km wide, with a coastline of 1,600 km and an area of 5,560 km^2 (Figure 4-1). The cool, humid climate is mainly influenced by continental air masses that are humidified and temperature-moderated by the surrounding ocean waters. January and July mean temperatures are −7°C and 18.7°C respectively with an annual mean precipitation of 1100 mm. The frost-free period varies from 100 to 160 days allowing for the cultivation of a wide variety of crops (Jiang et al., 2015).

A major physiographic unit in the Canadian Maritime provinces is the Maritime Plain which is almost entirely submerged by the Gulf of St. Lawrence. This lowland plain, including PEI, is underlain by late Paleozoic sedimentary rocks. In PEI, the bedrock formation contains a higher proportion of sandstone in the central and southeastern sections

of the island than in the western section, while the ratio of sandstone over claystone in the eastern section is less than those between the western and central sections (van de Poll, 1981; Jiang et al., 2015). Regionally, the bedrock is either flat lying or dipping to the east, northeast or north at an average of 1-3° with little structural deformation. The bedrock is overlain by a thin layer of coarse glacial deposits (0-10 m) derived from red sedimentary rocks. Soils derived from the glacial till are sandy and well drained (MacDougall et al., 1988). Although several soil types are identified based on variations of texture, the soils are considered relatively uniform across the island (e.g. Figure 4-2). The island's land surface is rolling; the western section of the island has a gentle relief with slopes up to 7%; the central and southeastern sections are more hilly, including slopes up to 14% with the highest point at an elevation of 139 m above sea level; the eastern section follows a relief lying between those of the western and central sections (MacDougall et al., 1988). The spatial variations of topography and soil are related to the spatial distribution of surficial geological materials of sandstone and claystone within the bedrock formation.

Based on the legacy soil survey report (MacDougall et al., 1988), the dominant soil order in PEI is Podzol (Podzol, WRB 2014; Spodosol, USDA) formed from the soft, red sandstone bedrock, and is sandy and mostly red in color, low in bases and nutrients, and acidic reaction (van de Poll, 1989). Additional soil orders such as Gleysol (Glesol, WRB 2014; Aqua-suborders, USDA), Luvisol (Luvisol, WRB 2014; Boralfs and Udalfs, USDA) and Brunisol (Cambisol, WRB 2014; Inceptisols, USDA) are also present. This study was conducted in the Maple Plains watershed, a well-defined watershed in a central area of PEI (Figure 4-1). The landform and soil associations of this watershed are common across the

island. There is a mixture of land uses, including crop production, riparian and woodlots. The surficial geological materials are nearly uniform except for some shallow phase bedrock soils found at higher elevation.

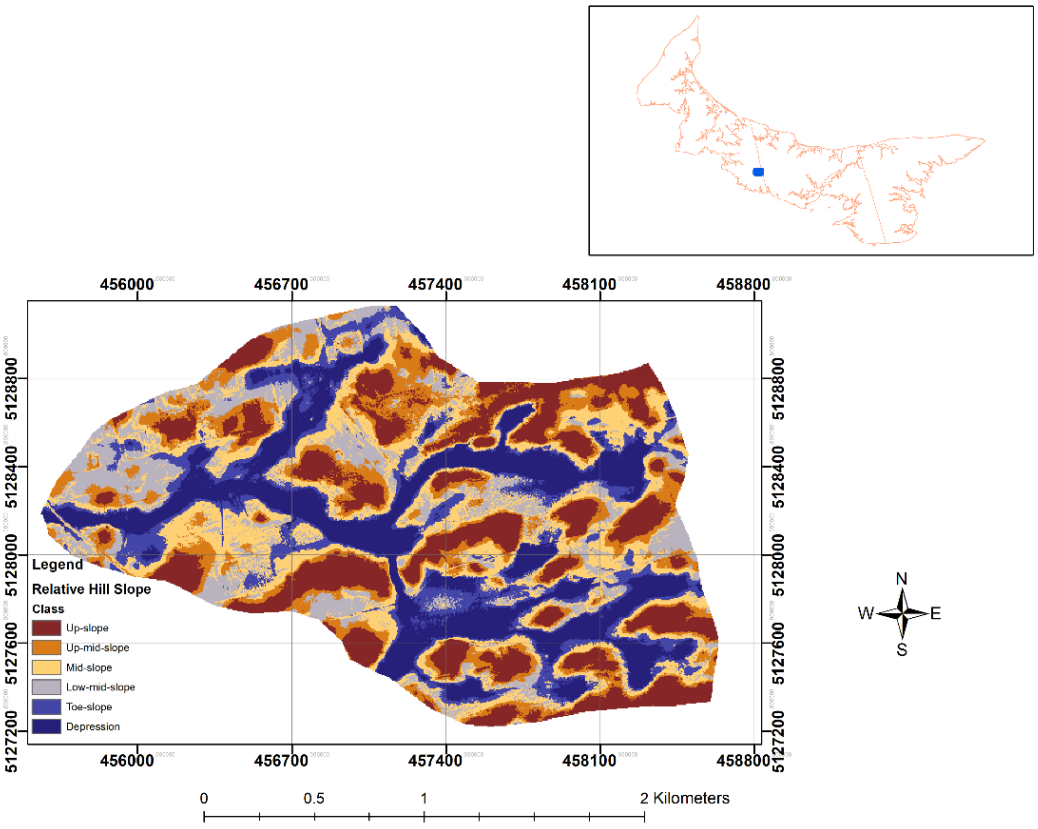


Figure 4-1 Maple Plains Watershed site location(main map) in Prince Edward Island (inset), Canada (DEM data source, “GIS data layers: geographic information for PEI”, http://www.gov.pe.ca/gis/, accessed May 7, 2020)

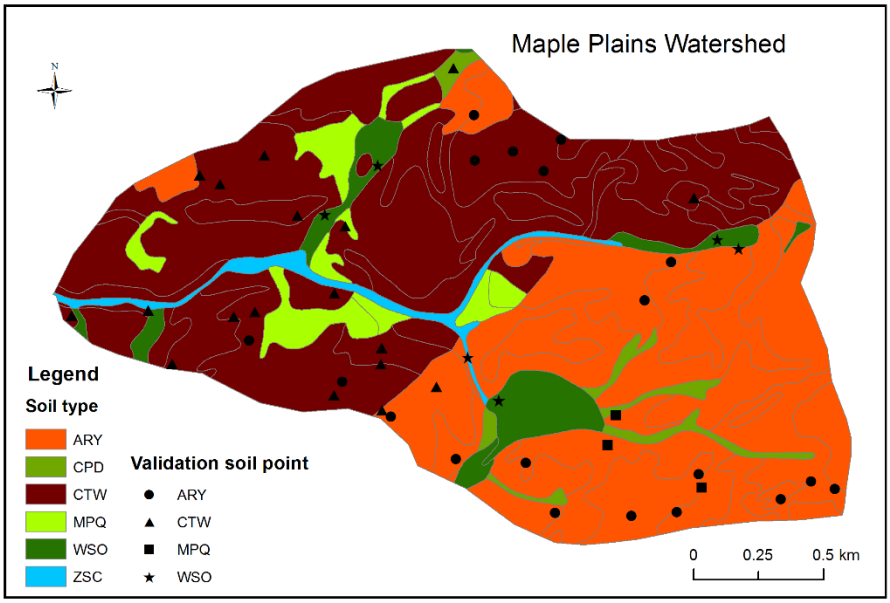


Figure 4-2 Legacy soil survey and validation sample locations at PEI study site. See Section 4.3 for soil type definitions (Soil survey data source, "GIS data layers: geographic information for PEI", http://www.gov.pe.ca/gis/, accessed May 7, 2020)

4.3 Data and method

4.3.1 Legacy soil survey

A 1:20,000 soil survey map along with associated attribute information was used for this study. For the entire province, the attribute table of the soil survey was only 80% complete, meaning that some of the surveyed polygons had not been linked to specific soil attributes. However, for the Maple Plains Watershed study area, the soil survey was complete. In the Canadian System of Soil Classification (Soil Classification Working Group, 1998), the soil types mapped from the legacy soil survey include Alberry (ARY) - Paralithic Orthic Humo-Ferric Podsol, Crapaud (CPD) - Gleyed Eluviated Eutric Brunisol, Charlottetown (CTW) - Orthic Humo-Ferric Podsol, Malpeque (MPQ) - Gleyed Eluviated Dystric Brunisol, Winsloe (WSO) - Orthic Gleysol, and Stream Complex (ZSC) - Wetland and Water. Note

that from here on, just the three-letter abbreviation is used to identify each of these soil types. In this study, surveyed soils with subtle differences within the same order were merged: gleysols CPD and MPQ were grouped to one soil type called MPQ, where small differences in pH were expected (Dystric pH < 5.5 vs. Eutric pH ≥ 5.5); both hydric soils WSO and ZSC were grouped to one called WSO. After the grouping, the soil series examined and mapped were CTW, ARY, MPQ and WSO.

3.2.2 Digital elevation model and landscape segmentation

The original DEM was acquired by the PEI government using an airborne Light Detection and Ranging (LIDAR) sensor ("GIS data layers: geographic information for PEI", http://www.gov.pe.ca/gis/, accessed May 7, 2020). In order to reduce unwanted "noise" from fine resolution LIDAR data, for the purpose of soil mapping, the 1 m Lidar DEM was further resampled to a 5 m pixel size for this study using the bilinear interpolation method. The coordinate system is WGS 1984 UTM 20N.

Multi-resolution topographic covariates except for the Topographic Index (TPI) were derived using the 5 m DEM and SAGA GIS (Conrad el al., 2015). The TPI was calculated using the Relief Analysis tool (Miller, 2013), an ArcGIS® toolbox designed based on Weiss (2001). More details about each of the topographic covariates are described below.

- Elevation: height above sea level extracted from DEM.
- Channel network base: channel network base level elevation.
- MRVBF: Multi-resolution valley bottom flatness aims to classify valley bottoms as flat and low areas (Gallant and Dowling, 2003).

- Slope: Topographic gradients were calculated and represented using percentages.
- Topographic position index (TPI) was calculated using a circular shape moving window approach. The inner radius of a ring-shaped moving window was 3 cells and the outer radius was 15 cells.
- Topographic wetness index (TWI) was calculated using a TWI routine from the SAGA tool kit.
- Valley depth: Valley depth was calculated as the difference between the elevation and an interpolated ridge elevation.
- Vertical distance to channel network: This covariate was derived by calculating the vertical distance to a channel network base.

4.3.2 Legacy soil survey data mining and training data

Regardless of the method(s) used in PSM, point soil data or soil samples either from legacy data or newly collected sources are essential inputs for statistical algorithms, training of machine learners and validation of predicted soil classes or properties. Statistically designed and sampled point data can represent the relationship between soil class/properties and environmental variables effectively, which in turn can produce more reliable maps. Poorly designed and unrepresentative samples (e.g. roadside or transect samples) are the major error source in PSM (Markert, 2007; Kidd et al., 2015).

Although we have hypothesized that any location within a single component soil survey polygon can be used to represent the single soil type of the polygon, different soil sampling locations within a generalized soil polygon may represent the central concept of the soil

development differently. For example, soil from one location may represent the mapped soil more closely than from another location within the mapped polygon. This may be more likely when point soil data from coarser or small scale soil surveys are extracted compared to that using finer or large scale versions. The four data mining or sampling methods used in this study are listed below. For each of the methods, the soil type at a sampling location was labelled as that of the polygon from the legacy soil survey (Figure 4-3).

- Simple random sampling: each location to be sampled was chosen randomly and had equal probability of being sampled.
- Stratified random sampling: the study area was first stratified into n (i.e. the number of locations to be sampled) sub-areas and then one sample was randomly selected within each sub-area.
- Area-weighted random sampling: the number of randomly sampled locations per soil polygon was determined by the percent of the area of the soil polygon of the entire study area.
- Conditioned Latin Hypercube Sampling (cLHS) and optimization: a model-based sampling design method that statistically provides an efficient way of sampling variables from their multivariate distributions. In this study DEM-derived slope and TWI were used as inputs for cLHS modeling.

Biswas and Zhang (2017) reviewed and studied several statistical and geometrical sampling designs. The recommendation is that a sampling design with a modeling approach is better than model-free ones; a sampling design should include both geospatial coverage

and issue-related geospatial variables or spatial features. A practical method that covers both geospatial coverage and associated special features is the cLHS approach (Minasny and McBratney, 2006). The cLHS approach is an advanced version of the original Latin Hypercube Sampling (LHS) method published by McKay et al. (1979). The core logic of the LHS is that in a given coverage of an area along with k variables, the cumulative distribution of each variable is divided into the number of intervals that match n number of samples to be taken with equal probability. A sample is randomly taken at each interval. There will be n number of samples for each variable. The selected n values for each of the variables are randomly matched. At the end of the iterations, n number of samples is selected for the whole distribution of variables. As an improved LHS method, the cLHS solutions can avoid generating sample points from locations where one or more variables do not have meaningful data.

The cLHS itself will distribute a predefined number of samples using the Latin Hypercube algorithms. There is no attempt to define the optimum number of samples to be taken. In order to find the optimized number of samples for a study area, cLHS is repeated based on a set of covariates with incremental sample sizes, e.g. from 10 to 500 with increments of 10. Normalized measurements of principal component similarity (Krzanowski, 1979), percentage of grid points within sample Principal Component (PC) space (convex hull) and Kullback-Leibler divergence (Kullback and Leibler, 1951) were used on the Y axis to plot against various sample size on X axis. An exponential decay function is then fitted to each of the measurement curves. The optimal sample sizes based on each of measurements are the sample numbers which are located at leveled-off points of the curves, or the points

which satisfy a user defined confidence level of 90% or 95% (Malone et al., 2019). For this study site with selected covariates at 5 m resolution, the average optimum sampling number based on the three measurements was 100 points at a 90% confidence level (Figure 4-3).

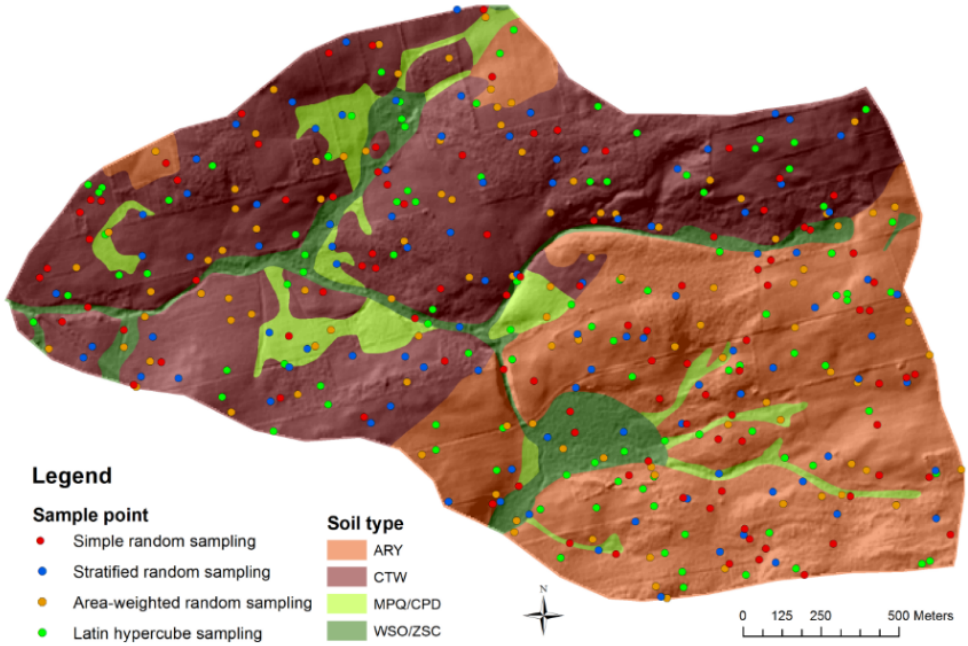


Figure 4-3 Data mining locations of the four studied sampling design methods (soil type data from http://sis.agr.gc.ca/cansis/nsdb/dss/v3/index.html, accessed May 7, 2020)

4.3.3 Validation data set

In this study, no legacy pedon data were available for validation. New point data were collected using the cLHS design. Considering the cost of field work, only an affordable number of sampling locations was planned. The cLHS sample locations were identified based on 30 m DEM derived LS-factor and slope gradients. Forty-seven of the 50 planned

points (ten times the number of dominant soils in the study area) were inspected and sampled in November, 2016. At each point, soil samples for bulk density and other properties such as soil particle sizes and soil organic carbon content, were taken from soil genetic horizons (top (e.g. Ap) and subsurface (e.g. Bf) horizons) using an Eijkelkamp soil sampling kit with a 5 cm diameter sampling tube. For each sampling point, the soil properties and field observations were used to determine the soil type (CTW, ARY, MPQ or WSO). A Trimble GeoXT 3.5G GPS receiver with horizontal accuracy of +/- 3 m was used to record the final sampling locations.

4.3.4 Soil type prediction using machine learning

Many inference and machine learning methods have been studied and used for PSM (Heung et al., 2017; Hengl and McMillian, 2019). In this study, RF, ANN, C5, and SVM with four training data sets were used to predict soil types. Customized machine learning procedures were written (examples provided in Appendix B), using R (R Core Team, 2018) and several R libraries for geospatial analysis/management and machine learning, including Rgdal (Bivand et al., 2019), Raster (Hijmans, 2015), and Caret (Kuhn et al. 2019). For the RF machine learning, training location samples were randomly divided into 70% and 30% portions as part of the internal boot strapping and cross validation use.

4.3.5 Accuracy assessment

The accuracy of the predicted soil maps was evaluated in three different ways. First, the overall accuracy was determined by the percentage of all the validation data being correctly allocated, which provided a general index relating to the accuracy of map predictions.

Next, accuracies of predicted soil types can be depicted by the calculation of producer's accuracies and user's accuracies. Producer's accuracy accounts for errors of omission and is the overall proportion of correctly classified ground truth soil class pixels. User's accuracy is a measure of map reliability and is the overall proportion of all pixels classified as a particular class that were correctly classified (Lillesand and Kiefer, 2000). As part of the categorical data accuracy assessment, Kappa values were calculated as overall agreement/disagreement assessments between the predicted and validation soil types. Kappa values ranged between 0 and 1, where 1 indicated full agreement and 0 indicated no agreement.

4.3.6 Classifier combination using majority voting

It is common for PSM projects to use multiple machine learning methods (i.e. an ensemble). This study used an ensemble of four machine learning classifiers: RF, C5, ANN and SVM. For each of the mapping grid cells, the output of the machine learning algorithms can be presented as a vector of numbers. The dimension of the vector is the number of the soil classes or types (Tulyakov et al., 2008). The task of this classifier ensemble is to find an optimum class or soil type for each mapped grid cell. Among many classifier ensemble methods, the simplest one is majority voting, which in this case chooses the most frequent class number among the predicted soil classes. To avoid ties, an odd number of predictions was used for each cell, by including equally-weighted predictions from the machine learners with the highest accuracy rankings for that location. Majority voting is the approach used in this study.

4.4 Results

4.4.1 Point soil training data mining

The number of soil data points generated using each of the sampling methods (Table 4-1) reflects the general soil class proportions for the study site. In the watershed, the areal ranking of the dominant soils was CTW > ARY > MPQ > WSO (Figure 4-3).

Table 4-1 Number of soil samples using four sampling methods, per soil type.

Sampling Method:	Soil type:				Total samples:
	ARY	CTW	MPQ	WSO	
Simple random	41	40	6	13	100
Stratified random	33	49	8	8	98
Area-weighted random	39	46	10	8	103
Latin Hypercube	46	35	8	11	100

4.4.2 Predicted soil type maps

Raster soil type maps (Figure 4-4) at 5 m grid sizes were produced for the 16 combinations of sampling methods and machine learning algorithms. All the machine learning predictions revealed similar soil distribution patterns compared to those presented via the legacy soil survey (Figure 4-2). However, the legacy soil survey can only present spatial soil information using areal polygons with crisp area class boundaries. The machine learning predicted maps explicitly map soil types pixel by pixel, therefore more variability can exist in a given area. The predicted dominant soils were CTW and ARY. ARY soil was found in the areas where the elevations are the highest in the study watershed. ARY soil is a shallower bedrock phase of the CTW. In machine learning predictions (Figure 4-4), ARY

soil was mapped as dominant in the northern/upper section of the predicted maps, in contrast to the legacy soil survey which showed more CTW soils identified in that section of the watershed (Figure 4-2). My field inspection and sampling agreed more with the predicted maps than with the legacy soil survey maps.

4.4.3 Accuracy assessment

The overall accuracy of the predictions ranged between 51% and 74% (Table 4-2). Among the machine learning procedures of RF, C50, ANN and SVM, RF performed the best with prediction accuracies above 70% across the data mining methods. RF with the simple random sampling method predicted soil types with the highest overall accuracy of 74%. Other machine learning methods also predicted soil types with higher than 50% accuracies. Among the studied sampling methods, simple random sampling had prediction accuracies of 74%, 70%, 67% and 65% when using RF, NN, SVM and C50 machine learning procedures, respectively. Table 4-3 lists the user's and producer's accuracy values for predicted soil types over each of the combinations of the machine learning algorithms and sampling methods. Overall, WSO soils were the most-accurately predicted according to producer's accuracy values. Again, the RF performed best among the machine learning approaches indicated by the higher average producer and user accuracies over all the training data sets. MPQ soil has such limited presence that the classified pixels were not sufficient for meaningful statistics, therefore there are zero values for MPQ soil in Table 4-3. As a gleyed soil in nature, MPQ is also blended into the WSO group soils.

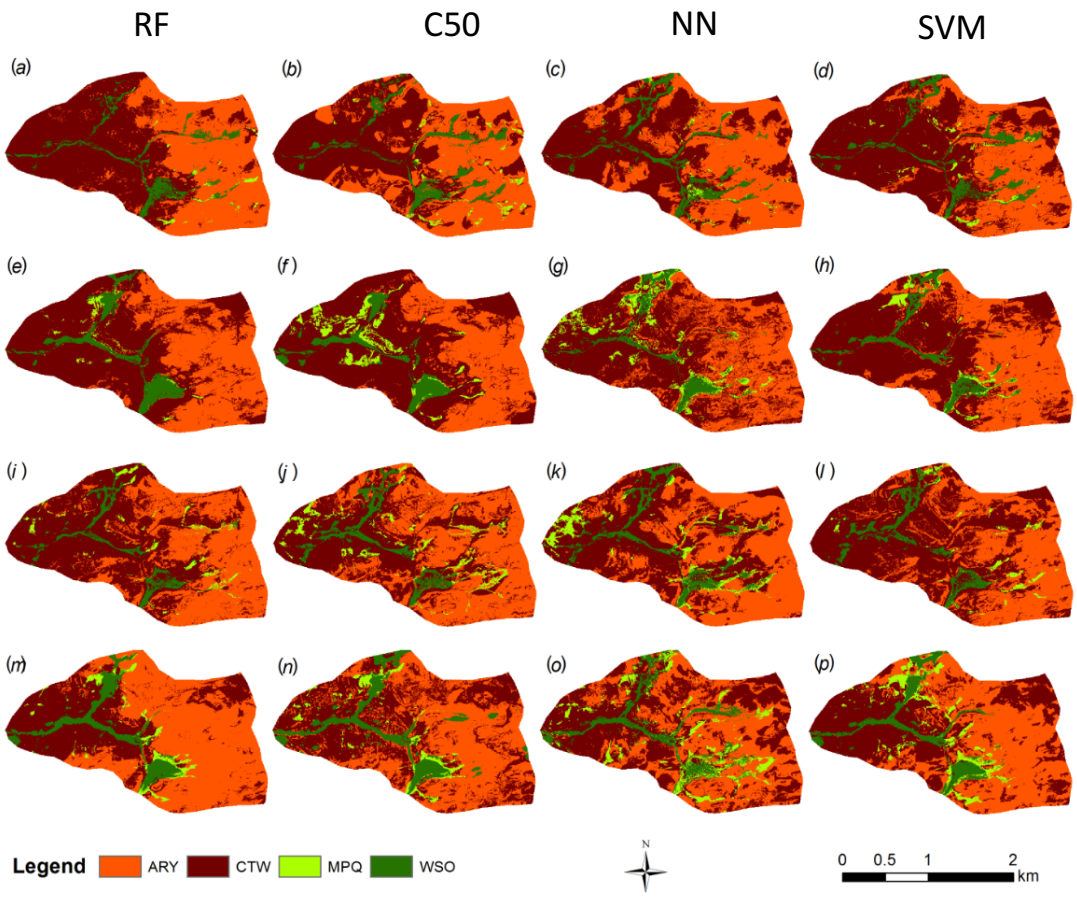


Figure 4-4 Predicted soil type maps with different sampling and machine learning methods. For each method from left to right (RF, C50, NN and SVM), maps from a to d show the predicted soil types using the fully random samples; maps from e to h were predicted using the stratified randomly selected samples; maps from i to l were predicted using the area-weighted random samples; and maps from m to p were produced using the cLHS based samples.

Table 4-2 Overall accuracy (OA) and Kappa coefficient between sampling methods and machine learning techniques

Sampling Method	RF		C50		NN		SVM	
	OA %	Kappa	OA %	Kappa	OA %	Kappa	OA %	Kappa
Simple random	74	0.59	65	0.46	70	0.51	67	0.48
Stratified random	70	0.51	52	0.24	67	0.48	57	0.30
Area-weighted random	74	0.58	.63	0.41	72	0.55	65	0.44
Latin Hypercube	70	0.51	54	0.27	57	0.34	67	0.48

4.4.4 Multi-machine learning and ensemble modeling

The prediction accuracies presented above were used to choose which combinations of methods would be applied in majority voting to determine the final soil type of that pixel for each grid (Table 4-4). For example, in Figure 4-5, map (*a*) is the majority map from all the training data collection methods under the RF machine learning only; map (*b*) is the majority map from the four machine learning methods using the area-weighted training data set; and map (*c*) is the majority map from the four machine learning methods using the simple random training data set. As shown in Table 4-5, all the final maps have higher than 70% accuracy. The RF machine learning based majority or final map has accuracy as high as 78%, as well has higher producer's and user's accuracies for the soil classes relative to the average accuracy values from all the RF results of the four differently mined training data sets.

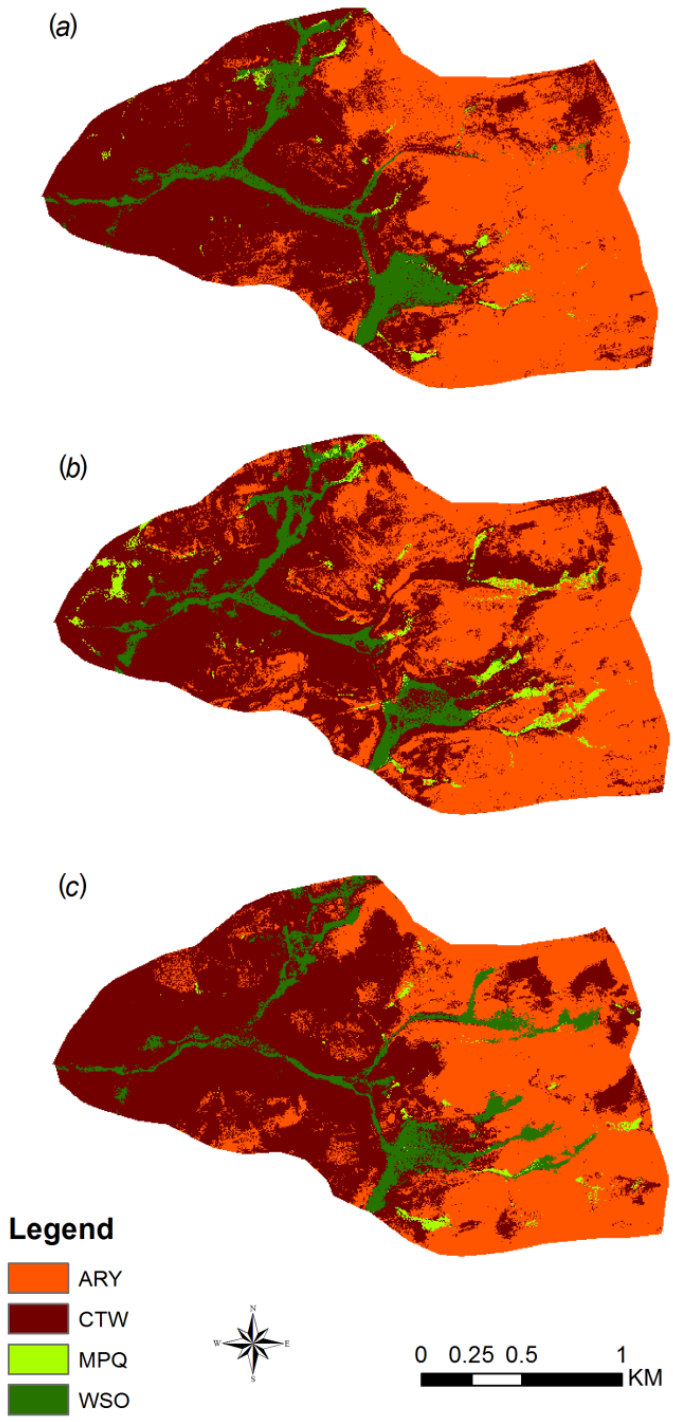


Figure 4-5 Optimized soil type maps using majority ensemble methods: a) the majority map from all the training data collection methods with the RF only; b) the majority map from the four machine learning methods with area-weighted training data set; c) the majority map from all the machine learning methods with simple random training data set. See Table 4-5 for accuracy assessments.

Table 4-3 User's and producer's accuracy of each predicted soil type

Sampling Method	Machine Learner	User's Accuracy			Producer's Accuracy		
		ARY	CTW	WSO	ARY	CTW	WSO
Simple random	RF	63.2	94.4	83.3	75	68	100
	C5.0	52.6	77.8	100	62.5	60.9	100
	NN	73.7	77.8	66.7	63.6	70	100
	SVM	52.6	94.4	66.7	71.4	63	100
Stratified random	RF	63.2	88.9	66.7	70.6	64	100
	C5.0	47.4	72.2	33.3	64.3	48.1	66.7
	NN	63.2	83.3	66.7	66.7	65.2	100
	SVM	52.6	72.2	50	52.6	54.2	100
Area-weighted random	RF	84.2	88.9	33.3	72.7	76.2	100
	C5.0	68.4	77.8	33.3	72.2	60.9	66.7
	NN	68.4	88.9	66.7	72.2	69.6	100
	SVM	57.9	94.4	33.3	68.8	63	100
Latin Hypercube	RF	84.2	66.7	66.7	61.5	80	80
	C5.0	63.2	50	66.7	48	56.2	80
	NN	68.4	55.6	50	65	62.5	60
	SVM	68.4	83.3	50	65	68.2	75

Table 4-4 Overall accuracy (OA), Kappa coefficient, producer's and user's accuracies for the majority maps

Averaging method*			ARY		CTW		MPQ		WSO	
	OA	Kappa	UA	PA	UA	PA	UA	PA	UA	PA
1	78	0.65	78.9	71.4	94.4	81.0	0	0	66.7	100
2	72	0.55	73.7	69.95	84.7	72.1	0	0	62.5	95
3	70	0.51	73.7	70.0	88.9	69.6	0	0	33.3	100
4	69	0.38	69.7	71.5	87.5	67.4	0	0	41.7	91.7
5	72	0.55	63.2	70.6	83.3	65.2	0	0	100	100
6	69	0.51	60.5	68.1	86.1	65.5	0	0	79.2	100

*Averaging method. 1—majority voting from all training data sets using RF; 2—average values from all training data sets using RF ; 3—majority voting from all machine learner using the area-weighted samples; 4—average values from all machine learner using the area-weighted samples; 5—majority voting from all the machine learner using the simple random samples; 6 average values from all the machine learner using the simple random samples.

Table 4-5 Overall accuracy (OA) and Kappa coefficient of mapped soil types using majority voting

Ensemble data source for Majority Voting	OA	Kappa
1 Soil maps from RF and four sampling methods	0.78	0.65
2 Soil maps from Area Weighted Sampling and four machine learners	0.70	0.51
3 Soil maps from Simple Random Sampling and four machine learners	0.72	0.55

4.5 Discussion

With detailed soil surveys at scales 1:20,000 or larger, many mapped soil polygons have single soil components, and we can achieve high prediction accuracies of soil types by extracting training data sets from any location in the sampled soil survey polygons. Despite the high overall accuracy obtained in this study, many detailed soil surveys with scales 1:50,000 or smaller overly generalize the relationship between soil and the environmental factors that influence soil development. Randomly locating a data mining point in a soil polygon may miss the locations where they can represent the mapped soil accurately. Although the fully random sampling method provided effective point soil information for accurate soil type machine learning predictions in mapping the Maple Plains Watershed soils, other sampling methods or data mining approaches such as the cLHS may be used when a fully random sampling cannot extract more representative soil information via legacy soil survey. In the case of extracting training data sets from multi-component soil survey polygons, Odgers et al. (2014) developed the DSMART tool which finds sampling locations based on soil genesis information embedded within soil surveys. However, there are only a few application examples of the DSMART tool (Zeraatpisheh et al., 2019). The

preliminary application of DSMART during the FAO soil carbon mapping project (FAO, "Global soil organic carbon map (GSOC)." http://www.fao.org/global-soil-partnership/pillars-action/4-information-and-data-new/global-soil-organic-carbon-gsoc-map/en/, accessed May 7, 2020) contributed to producing a national extent soil carbon map for Canada, but no validation has been conducted. DSMART, as its name indicates, it is a tool mainly for disaggregating soil surveys with multi component soil polygons. More research is needed in this field.

Among the machine learning algorithms, RF showed the best performance in terms of prediction accuracy in this study (70% - 74% overall accuracy) as shown in Table 4-2. One of the reasons that a RF machine learner outperforms the others is that the RF is an ensemble and multiple decision tree classifier itself, using internal cross-validation through bootstrapping and out-of-bag procedures. By contrast, C5 had the lowest overall accuracy among the machine learners, and it doesn't include cross-validation and bootstrapping procedures. The application of ANN and SVM require customized parameters and longer processing times, which may be more problematic for larger areas (e.g. Canada wide) and finer resolution (e.g. 50 m grid size) mapping across the nation. Each of the selected machine learning methods can perform better in some mapping areas and worse in others. To further improve the accuracy of results from a single method, the outputs of all machine learning approaches in this study were incorporated in an ensemble framework. Among the ensemble algorithms and frameworks, simple majority logic was used for extracting and optimizing the final soil type map(s) (Figure 4-5), following the example of Malone et al. (2014), which concluded that ensemble procedures produced higher accuracy maps than

any of the studied individual predictions in mapping area in Queensland, Australia. With this ensemble majority logic, the overall prediction accuracy was further increased as high as 78% (Table 4-5). This prediction accuracy is higher than those reported either at global scale (overall cross validation accuracy 61%) (Hengl et al., 2017) or at a watershed scale in another Canadian context (61-68%) (Heung et al., 2017). This higher accuracy may be due in part to the relatively small project area and limited diversity of soil types in this case.

PSM using machine learning methods has been widely used (Hengl and McMillian, 2019). PSM should be conducted according to specific objectives and needs. In the context of mapping objectives, the prediction accuracy could be also improved by grouping related soil types before running the prediction algorithms or by collecting sufficient soil point training data. For example, an objective to map out well drained and poorly drained soils can be more accurately realized than an objective to map out within field variability for variable rate fertilization; predicting soil types at coarser thematic resolution (fewer categories) has less uncertainty.

In this chapter, soil series developed from consistent parent materials and taxonomic names were grouped. The decision to group soil series is well justified. In Canada, the soil series names are related to geographic place names, but despite the different names, many soil series have similar soil properties. Treating those reported soil names differently during PSM will further propagate the confusion inherited from the legacy soil surveys. Legacy soil survey-reported soil types need to be grouped according to taxonomic similarities to

meet PSM objectives that include providing information related to soil functions (e.g. agriculture, water management, etc.). For example, CTW and MPQ soils had similar sub-group names in this study and they were merged into one group. In the context of the Canadian Soil Classification System, soil survey reported soil names should be evaluated and grouped based on the soil association and catena definitions. A soil association includes member soils which have similar soil parent materials. A soil catena includes soil association members which are orderly connected from higher to lower landscape positions. Soil name merging and grouping is especially needed for larger (even nationwide) PSM procedures (Scull et al., 2003).

4.6 Conclusions

Legacy soil survey data can be used to extract the needed point soil information for machine learning algorithms. Although we conclude that simple random sampling among those single component soil survey polygons can lead to higher prediction accuracy of soil types in this study, the extension of this conclusion should reference the scale of soil survey to be used. In other contexts, other sampling methods may lead to better outcomes of machine learning. The RF, C5 decision tree, ANN and SVM machine learning algorithms all performed well, with greater than 50% overall accuracy of the predicted soil types, but RF performed the best with prediction accuracy above 70% across the sampling methods. The best performing sampling/data mining method was simple random sampling, which had prediction accuracy of 74%, 70%, 67% and 65% from using RF, NN, SVM and C50 machine learning procedures, respectively. A majority appearance-based ensemble optimization of the predictions from the machine learning brings the overall accuracy up

to 78%. In this study site, with the overall high accuracies of predicted soil types using the four different sampling methods, I conclude that anywhere within a single soil component soil survey polygon can be sampled for PSM.

Chapter 5: **Methods of soil property inference using a predicted soil class map**

5.1 Introduction

Soil classes and properties are typical categorical and continuous dependent variables in PSM. In the last two decades, most PSM studies have aimed to infer soil classes (Yang et al., 2010; Liu et al., 2016). Efforts to map soil properties using predictive soil mapping methods have recently been made (McBratney et al., 2003; Zhang et la., 2017). Estimating soil and physical properties for areas without ground sampling is especially important for environmental management, modeling and agricultural operations (Grunwald et al., 2011; Thompson et al., 2012; Silva et al., 2016). Not all soil properties need to be digitally mapped, because some key soil (physical) properties can be used to derive other soil properties using pedo-transfer functions (Van Looy et al., 2017).

Soil physical and chemical properties can either be estimated directly using sufficient training or point soil information and geo-statistic methods (Heuvelink and Webster, 2001; Hengl et al., 2017) or be inferred indirectly via mapped soil classes along with representative soil properties and probabilities of mapped soils at a location (Zhu et al., 2010a; Adhikari et al., 2013; Akumu et al., 2015; Viscarra Rossel and Bui, 2016; Camera et al., 2017). Pedometrics or statistically-based methods often require dense point soil information (Henderson et al., 2005; Bui et al., 2006, 2009; Cianfrani et al., 2018). However point data are either insufficient or out of date in most places in Canada. Even in

locations where there are sufficient and statistically distributed point soil data, different PSM methods used to map soil properties still have relative advantages and disadvantages (Beguin et al., 2017). Other challenges include how to evaluate prediction uncertainties (Arrouays et al., 2017) and accuracies. In summary, there is a need to further study and develop a suite of holistic, adaptive and practical approaches for soil property mapping in Canada.

The main objective of this chapter is to study and develop methods of soil property mapping while following the latest predictive soil mapping protocols from global soil mapping initiatives, as outlined in Heuvelink (2014). The aim will be to predict soil properties indirectly via predicted soil types plus directly sampled point data.

5.2 Methods

5.2.1 Study site and legacy soil survey data and soil attribute pre-processing

Using the same PEI Maple Plains watershed described in Chapter 4, soil series examined include Alberry (ARY) - Paralithic Orthic Humo-Ferric Podsol, Crapaud (CPD) - Gleyed Eluviated Eutric Brunisol, Charlottetown (CTW) - Orthic Humo-Ferric Podsol, Malpeque(MPQ) - Gleyed Eluviated Dystric Brunisol, Winsloe (WSO) - Orthic Gleysol, and Stream Complex(ZSC) - Wetland and Water. Among the surveyed soil series, soils with limited areal coverage and similar taxonomic great groups (Soil Classification Working Group, 1998) were grouped. CPD and MPQ were grouped to one soil type called MPQ. The initial pH differences between PD and MPQ soils noted when first surveyed in 1981 have

diminished, likely due to recent agricultural land use. WSO and ZSC were grouped to one called WSO as both soils were hydric. After grouping, the soil types CTW, ARY, MPQ and WSO, along with the probabilities of finding each of the grouped soils, were mapped using majority voting from all training data sets, using RF as described in the previous chapter (Figure 4-5a). Those probability data sets corresponding to each of the grouped soils series were used to calculate weighted mean property values of each mapped pixel. Representative soil properties such as soil bulk density (BD) (g/cm^3), soil texture, and soil pH were reported in the legacy soil survey reports (MacDougall et al., 1988). Those properties were measured at each of soil genetic horizons but not at a uniform depth. Since these soil properties tend to vary with depth (Ponce-Hernandez et al. 1986) and there is a need to map soil properties at uniform depth (Arrouays et al., 2014), to calculate the weighted mean soil properties, a spline-based interpolation tool was used to estimate the soil properties measured by genetic horizons to a uniform depth (Bishop et al., 1999; https://github.com/cran/GSIF/blob/master/R/mpspline.R, accessed May 7, 2020). Table 5-1 provides examples of how properties such as BD were reported in the legacy soil survey report of 1988. It shows how those soil genesis horizon values were interpolated to the uniform soil depths (i.e. 0, 10, 30, 60, 100 m) using the spline tool. Similarly, the BDs of the other soils involved in this study were processed and are summarized along with newly sampled properties in Table 5-2 (see Chapter 4 and next section for more details). The mean BD value for the major soil groups at uniform soil depths (hereafter referred to as representative values) listed in Table 5-2 are used for further soil property mapping in this chapter.

Table 5-1 Survey soil bulk density (BD, g/cm^3) of Alberry (ARY) and Charlottetown (CTW) soils in the Maple Plains Watershed of Prince Edward Island (MacDougall et al., 1988)

Soil code*	Upper depth (cm)	Lower depth (cm)	BD (g/cm^3)
PEARYD A	0	20	1.1
PEARYD A	20	33	1.3
PEARYD A	33	53	1.5
ARYD A	53	84	1.5
PEARYD A	84	100	1.6
PECTWD A	0	20	1.3
PECTWD A	20	30	1.1
PECTWD A	30	60	1.3
PECTWD A	60	100	1.9

*****Upper and lower depths refer to the extent of the genetic horizon sampled. From the PEI soil survey report and NSDB soil naming convention, the soil codes are named by concatenating the Province Code + Soil Series Name + Modifier + Land Use (MacDougall et al., 1988). For examples, PECTWD A means PEI soil CTW (Charlottetown) has D slope class (5% to 9%) and land use type of A(agriculture)**.**

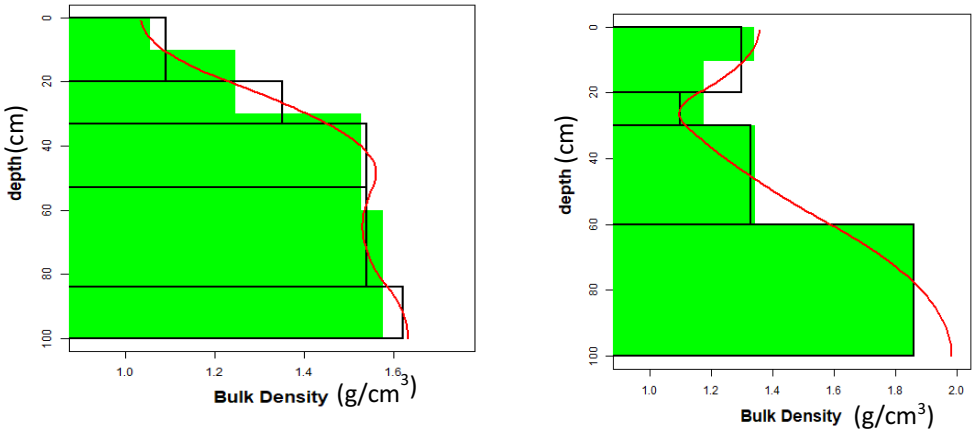


Figure 5-1 Spline curves of soil bulk density for Alberry (left) and Charlottetown (right) soil. The black boxes represent the observed data; the red curves are the continuous splines; the green boxes are spline averages.

Table 5-2 Interpolated soil bulk density of 0 – 10 cm depth (a recognized best practice in the PSM community) for both legacy and newly sampled representative soils (detailed in Chapter 3).

Soil_ID	Soil	Survey reported BD(g/cm^3)	Newly sampled BD(g/cm3)
1	ARY	1.1	0.8
2	CTW	1.3	1.0
3	CPD/MPQ	1.4	0.9
4	WSO	1.6	0.6

5.2.2 Soil sampling design for soil property mapping and validation

Initially, in order to validate mapped soil series, 47 locations in the Maple Plains Watershed were sampled, out of a desired distribution of 50 point based on the cLHS design described in Chapter 4 (3 sites provided to be inaccessible). Soil properties such as soil BD (g/cm^3), soil texture (percent of sand, silt and clay), and soil organic carbon were analyzed from two depths at each location. With the focus of this chapter on method development, only

soil BD data are used to demonstrate how soil property maps could be produced or renewed. Soil BD was determined as the oven dry weight of the soil divided by sampled volume (63 cm^3). Air dried soil samples were oven dried at 105°C for up to 24 hours.

Point to point validation was conducted based on the goodness of fit between the observed and predicted BD values. In this study, the goodness of fit was evaluated using coefficient of determination (r^2), concordance correlation coefficient (ϱ_c) and root mean square error (RMSE) (Equation 5-1, 5-2, and 5-3) (Hastie et al., 2009).

$$r = \frac{\sum_{i=1}^{n}(obs_i - \overline{obs})(pred_i - \overline{pred})}{\sqrt[2]{\sum_{i=1}^{n}(obs_i - \overline{obs})^2}\sqrt[2]{\sum_{i=1}^{n}(pred_i - \overline{pred})^2}} \qquad \text{Equation 5-1}$$

$$\rho_c = \frac{2\rho\sigma_{pred}\sigma_{obs}}{\sigma_{pred}^2 + \sigma_{obs}^2 + (\mu_{pred} - \mu_{obs})^2} \qquad \text{Equation 5-2}$$

where

$\rho\sigma_{pred}\ \sigma_{obs}$ is covariance between the predicted and observed values. σ_{pred} and σ_{obs} are variance of predicted and observed values respectively. μ_{pred} and μ_{obs} are the population means.

$$RMSE = \sqrt[2]{\left(\frac{\sum_{i=1}^{n}(obs_i - pred_i)^2}{n}\right)} \qquad \text{Equation 5-3}$$

5.2.3 Methods of soil property inference

In this chapter, it is assumed that any changes to soil classes and their spatial distribution across the landscapes of a mapping area between the time when they were originally surveyed and the present are negligible. The overall workflow of using mapped soil series

to indirectly infer soil properties is illustrated in Figure 5-2. Two specific implementations of the workflow were studied.

Method 1: Using Survey Reported Pedon Data (USRPD): Representative values of a soil property for each of the mapped soil series at a given depth were extracted from the soil surveys. The predicted value of a soil property for each mapped pixel is the weighted average of the representative soil attribute values where the weights are the probabilities of occurrences of the relevant soil series produced for each pixel using the RF method.

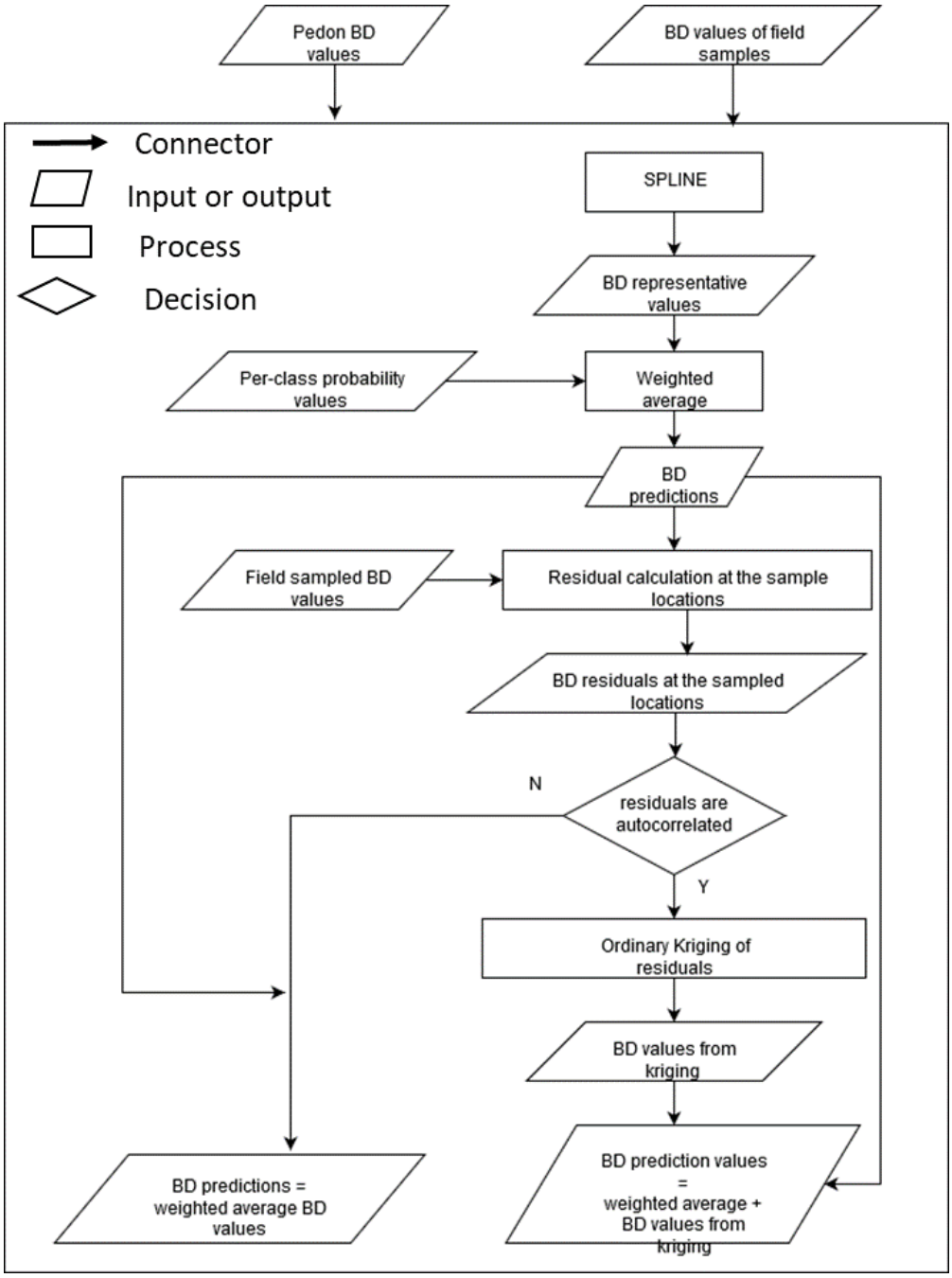


Figure 5-2 Workflow of soil property mapping using predicted soil series, probabilities and Spline interpolated soil point data

For a pixel, the weighted average of a predicted property is calculated as:

$$\text{Predicted} = \sum_{k=1}^{n} p(k) * r(k) \qquad \text{Equation 5-4}$$

where

$p(k)$: probability for soil k, k=1, 2, … n

$r(k)$: representative soil property value for soil k.

For example, with the interpolated representative BD data at 0-10 cm, for this study site:

Predicted = p[1] * 1.053 + p[2] * 1.339 + p[3] * 1.366 + p[4] * 1.608

where the representative BD values of 0-10cm depth are interpolated using spline model for each of the soils named ARY, CPD/MPQ, CWTD, and WSO representatively.

If there are legacy pedon point data available from a mapped area, the soil properties of those legacy pedons could be used for validation and residual mapping. In the Maple Plains watershed, no legacy pedon data are available. The soil survey properties for the mapped soil series were instead based on similar soil series properties surrounding this watershed. Newly surveyed locations within the watershed were inspected and sampled for BD as part of this study. The residuals or differences between the weighted predicted BD and newly measured BD were used for residual kriging (Equation 5-5).

$$\text{Residuals} = \text{observed} - \text{predicted} \qquad \text{Equation 5-5}$$

Ordinary Kriging (OK) was then conducted on the residuals. If the residuals have spatial structure, as determined using a variogram, the mapped residuals would be location-specific, quantifiable and auto-correlated (Minasny and McBratney, 2016), therefore they are added to the predicted soil property map (see workflow Figure 5-2). The uncertainty of

the final soil property map was captured with the standard errors of residuals produced during the residual kriging. At each of the mapped pixels, the standard error of the residuals was calculated with the overall mean of residual and specific residual value of the pixel. For each sampling and machine learning method, there was one specific residual value for each pixel and one over all standard error of the residuals across the mapping extent. The weighted standard error was further used to compute the upper and lower limit of prediction uncertainty which was calculated with using residual mean plus or minus the margin of error at 95% confidence interval.

Method 2: Using Field Sampled Pedon Data (UFSPD). Unlike Method 1, for each sampling and machine learning method, there was one specific residual value for each pixel and one overall standard error of the residuals across the mapping extent. The weighted standard error is further used to compute the upper and lower limit of prediction uncertainty at each pixel, which was calculated using the residual mean plus or minus the margin of error at 95% confidence interval (for more details about how the residuals were handled, please refer to http://spatial-analyst.net/ILWIS/htm/ilwisapp/kriging_algorithm.htm, accessed May 7, 2020).

5.3 Results

5.3.1 Mapped soil bulk density with soil survey reported and newly sampled point data

For soil BD of 0-10 cm depth, the predicted soil BD maps (Figure 5-3, a and b) were calculated using the values from Table 5-2 along with the soil series map (Figure 4-4, d)

and soil series probability maps (as illustrated in Figure 5-4). Visually, the predicted soil BD map shows similar spatial patterns to the mapped soil series (Figure 5-3, c and d). For the depth between 0 and 10 cm, the higher BD value of the WSO series pixels, which are gleyed soils developed on recent fluvial materials, agrees with those reported in the survey. The predicted BD map is continuous rather than an area class as presented in legacy soil surveys. The initial BD maps were used for further validation and residual treatment in Section 5.3.2.

5.3.2 Point to point validation, residual kriging and mapping uncertainties

At each of those sampled pixel locations using the cLHS design, the difference between the averaged survey reported and newly-measured BD values includes temporal and spatial changes of the soils overall. The differences or residuals from those sampled locations were tested for distribution normality using the Anderson-Darling test; no data transformation was required on these residual data ($p=0.229$).

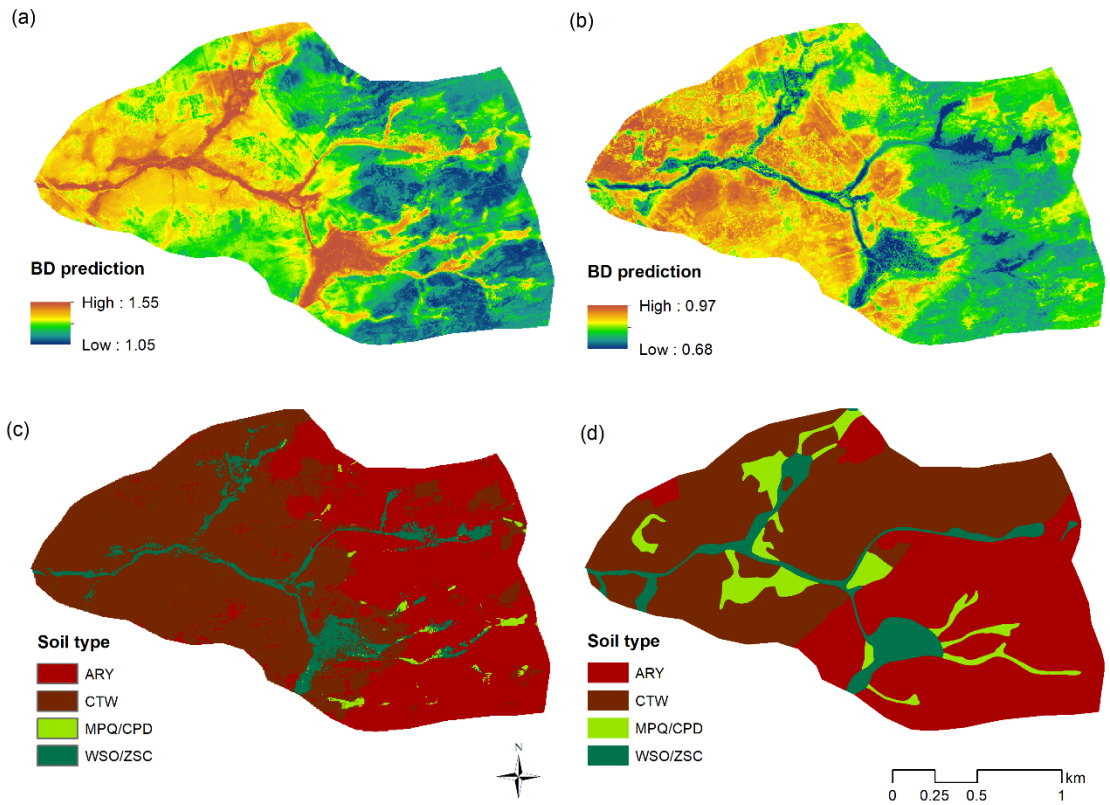


Figure 5-3 Predicted soil bulk density (g/cm^3) maps from (a) USRPD and (b) UFSPD methods; soil series maps predicted using (c) random forest and randomly sampled training data, and (d) soil survey maps from CanSIS (http://sis.agr.gc.ca/cansis/nsdb/dss/v3/index.html, accessed May 7, 2020).

The residuals were used to calculate the variogram and to interpolate across the mapping area using OK (Figure 5-5 and Figure 5-6). Since both the variogram and OK map of the residuals indicated autocorrelation of the residuals, the interpolated residuals were added back to the weighted map of BD. Using the standard deviation of the overall residuals, uncertainty of the predicted BD was represented with 90% range of the predictions (Figure 5-7, d, e, f). Since the regression kriging BD map had a normal distribution, 1.65 and -1.65 multiplied by the standard deviation were the upper and lower limit of the 90% interval range of the prediction respectively. The final soil maps using legacy and newly sampled

data sources were the ones with the added residual map in this study (Figure 5-7a and 5-7b). The BD values of the validation points were used for overall validation by computing the coefficient of determination (R^2), concordance (ϱc), and root mean square error (RSME) (Table 5-3).

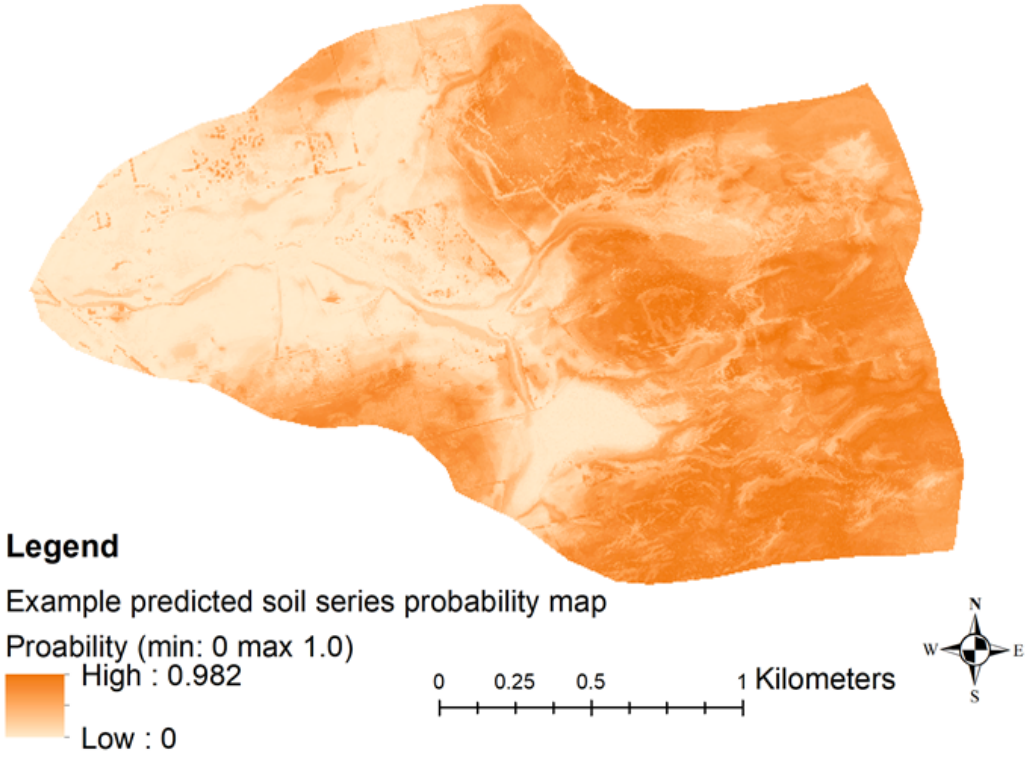


Figure 5-4 An example predicted soil series (ARY - Paralithic Orthic Humo-Ferric Podsol) probability map produced during soil class prediction in Chapter 4

Table 5-3 Accuracies of final predicted BD maps with two point data sources

Data source of BD map	R^2	ϱ_c	RMSE
Soil survey reported	0.92	0.88	0.089
Newly sampled points	0.90	0.76	0.126

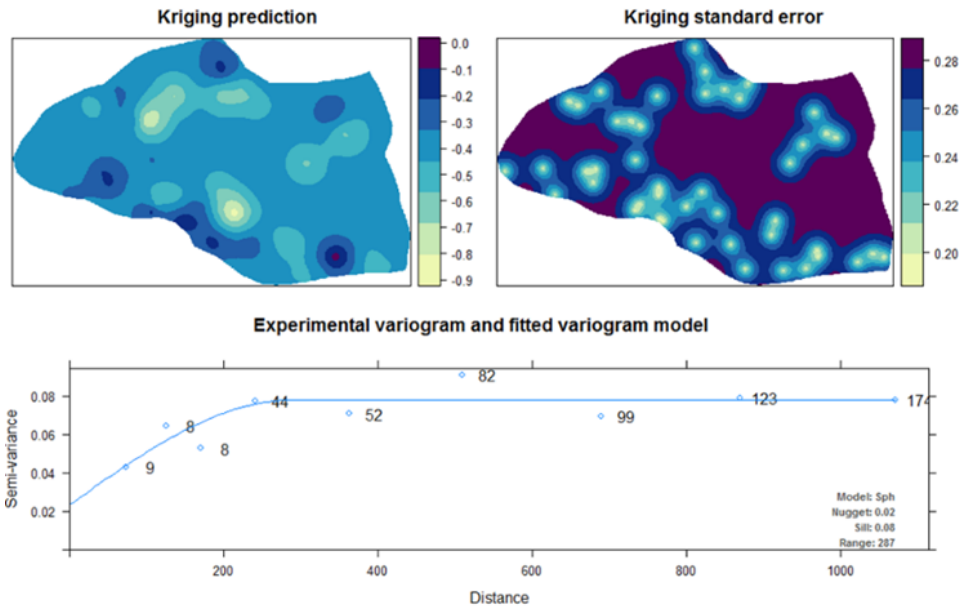


Figure 5-5 Fitted variogram models and OK maps of the residuals of the method 1 using survey reported pedon data

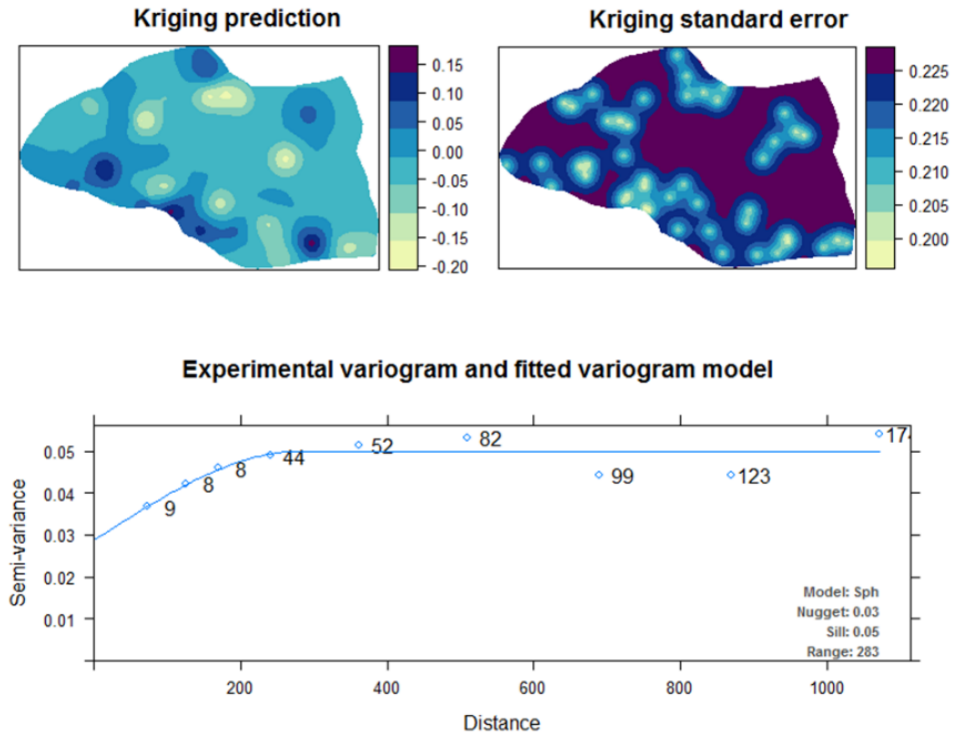


Figure 5-6 Fitted variogram models and OK maps of the residuals of the method 2 using newly surveyed pedon data

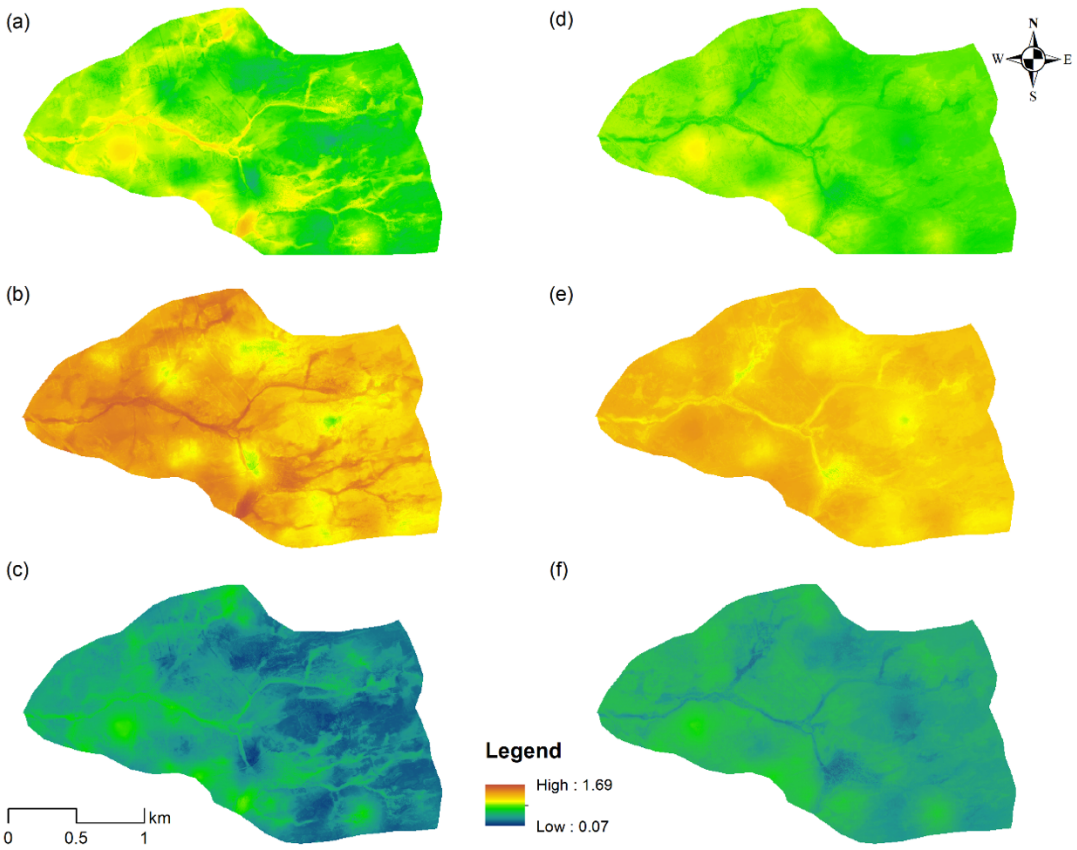


Figure 5-7 Mapped soil bulk density (g/cm^3) with soil survey reported (USRPD) versus newly sampled point data in the field (UFSPD):

(a) —prediction with USRPD method
(b)—upper interval of the prediction using USRPD
(c)—lower intervals of the prediction using USRPD
(d) —prediction with UFSPD method
(e)— upper interval of the prediction using UFSPD
(f)—lower intervals of the prediction using UFSPD

5.4 Discussion

Ideally there would be legacy point or pedon data available for validation and residual evaluation of soil property maps created using PSM. However, some soil series' properties may change slowly over time while other soil properties such as soil carbon content may change much faster due to land use conversion (Judith et al., 2017) and agricultural management measures such as liming and fertilization. For example, liming practices in agricultural fields can change soil pH values drastically (Carter and Richard, 2000). As a result, PSM methods to digitally map dynamic soil properties using legacy data alone may be questionable.

In this study, no legacy pedon data were available for validation of BD maps. Instead, 0-10 cm BD for the mapped soil series was based on similar soil series' properties surrounding the Maple Plains Watershed in order to carry out the USRPD workflow (i.e. using survey reported pedon data). This decision was justified on the basis of several studies which had some success in using legacy soil data for PSM outside of agriculture areas in Canada (Beguin et al., 2017; Bulmer et al., 2016). New point data were also collected using the cLHS design for the UFSPD workflow (i.e. using field sampled pedon data). Considering the cost of field work, affordability limited the number of sampling locations across the watershed. Having 50 points across four major soil series was sufficient for statistical use. The optimized cLHS method is recommended for any future sampling campaign based on a comprehensive review and research by Biswas and Zhang (2018). With the optimized sampling data, the prediction accuracy using newly sampled point data can be improved because the sampling locations are statistically calculated by

including soil development covariates (Biswas and Zhang, 2018).

Instead of sampling all soil properties, a number of soil properties could be inferred from other sampled property values along with selected co-variables. For example, soil water holding capacity can be further derived using predicted soil texture and bulk density values (Amirian-Chakan, et al., 2019). This methodology can be employed even though different soil properties such as pH, coarse fragments, and organic carbon content are not normally distributed spatially (Hengl et al., 2017). Machine learning algorithms such as RF can treat the nonlinear relationships better than geo-statistical methods such as ordinary kriging (Hengl and MacMillan, 2019). However, if training data sets are collected mainly by considering soil genesis factors, soil properties may be better predicted indirectly via predicted soil types. However, the mapping extent and scale are important for this decision. For example, for a relatively small hill slope or field, soil property mapping is more accurate and feasible with the UFSPD or USRPD workflows, while nationwide PSM of soil properties may be best carried out indirectly using soil types. For example, a soil carbon content map could be indirectly predicted across Canada by mining an existing wetland/peatland map at 1:1,000,000 scale (Tarnocai et al., 2011).

In this study, BD values from the same soil type were averaged and adjusted with predicted soil type probability values (see UFSPD and USRPD workflows). The Maple Plains Watershed only included four main soil types and was a relatively small area. For a large mapping area with a wider range of soil property variations, this indirect way of predicting soil properties should be further studied. One of the suggested solutions is to use the soil

similarity concept (Zhu, 1997) to find representative soil properties values of grouped soils. In a countywide or national soil survey, many reported soils are taxonomically similar although the local soil names are different. That is why grouping of some of the reported soils is needed. Grouping of soils using similarity indices for indirectly-predicted soil properties needs to be further studied.

In this study, there was a systematic difference between the BD values determined using the USRPD and UFSPD workflows. The reasons for the systematic difference are not known, but the spatial pattern and trend of BD of the mapped watershed were consistent. Some potential reasons for the differences include that the BD was measured and then interpolated from samples taken from different depths within sampling depth ranges, and that the legacy soil pedon data were reported from similar soil pedons outside of this studied watershed. Although the two workflow methods produced different BD values, upper and lower prediction intervals overlapped.

A clear goal for PSM-produced soil type and property data is to support geospatially enabled ecosystem models. For example, the recently released Soil Water Assessment Tool (SWAT) Plus model needs both soil types and soil property data such as soil particle size (i.e. sand%, silt% and clay%) at various depths organized by soil genesis horizons (Arnold et al., 2018). In order for the workflows presented here to achieve this goal, PSM should be conducted with understanding of and information on soil genesis of mapped soils. PSM is more than just a statistical and computing issue.

Chapter 6: **Microwave remote sensing and soil spatial pattern**

6.1 Introduction

Predictive soil mapping, including the mapping of soil classes and soil properties, involves the creation of new raster-based soil attribute datasets from existing soil and environmental covariate data. However, the expression and measurements of the relationships between soils and commonly used environmental co-variables are location- and scale-dependent, and across different locations, environmental covariates may not be uniformly available. Predictive soil mapping methods for soil type and soil properties need to be contextual and adaptive while observing the CLORPT soil development model.

In Canada, level and gently undulating landscapes are common. Generally, in these landscapes, the challenge of PSM is that the easy-to-observe factors such as topographic and vegetation conditions are not as effective for predicting soil types and properties as in places where topography (and land uses) are more diverse (Mendonca Santos et al., 2000; Iqbal et al., 2005). Other co-variables need to be considered such as remotely-sensed data (Odeh and McBratney, 2000; McBratney et al. 2003; Liu et al., 2012). Multi-spectral optical and time series remotely-sensed data have been used in PSM studies in China and Canada (Zhu et al., 2009; Liu et al., 2012). The studies are based on the hypothesis that the amplified differences of sensed spectral signatures before and after major rainfall events are related to spatial variation of soils and their properties. However, acquiring good quality optical remotely-sensed data immediately after major precipitation events can be difficult due to cloud cover or lengthy satellite revisit time. In contrast, microwave remotely-sensed

data can be collected in all cloud conditions (Zribi and Dechambre, 2002). There have been many applications of remotely-sensed data including microwave data for thematic mapping such as wetland extent, flooding, wetland classification and water cover fractions (Guo et al., 2017), but there have been fewer examples for operational PSM of wetland/peatland properties and extent (Manisny et al., 2019). Both passive and active microwave data have been used to map soil moisture over the last four decades (Karthikeyan et al., 2017). There are few examples using microwave data for operational PSM (Chang and Islam, 2000; Niang, 2008; Geng et al., 2010a), however, a number of studies have demonstrated the feasibility of combining important environmental variables such as topographic data with remotely-sensed microwave data for mapping (peatland) soils in Canada (Millard and Richardson, 2013; Li and Chen, 2005). With the recent availability of Sentinel-1 and Sentinel-2 data, the combination of optical and radar remotely-sensed data has improved soil moisture retrieval greatly (Gao et al., 2017; Hajj et al., 2017). In Canada the multi-polarization C-band RADARSAT-2 satellite system has been operational since 2009. RADARSAT-2 derived soil moisture distribution across landscapes is correlated with soil type and soil physical properties (Merzouki et al., 2011). The recently launched RADARSAT Constellation Mission (RCM) will increase the satellite revisiting time to almost daily. This makes time series data acquisition before and after a major rainfall event even more achievable. In addition to multi-polarization C-band microwave data from the RADARSAT-2 satellite platform, longer wavelength (e.g. L-band) synthetic aperture radar data, such as that from the Advanced Land Observation Satellite (ALOS), may be more suitable for mapping soil physical properties and spatial patterns (Baghdadi et al., 2008; Karthikeyan et al., 2017).

In this study, I hypothesize that the response surfaces of time series microwave data before and after a major rainfall event are correlated with soil parent materials/surficial geological materials and the associated soil classes. As one of the most important environmental covariates, soil parent material or surficial geological material data is often generalized or mapped at coarse scales in Canada. If time series remotely sensed data can detect the spatial pattern of the materials, finer resolution PSM in places where parent materials data are lacking becomes more feasible. The objectives of this chapter are twofold: 1) to evaluate the correlation between microwave data and soil parent materials/surficial geological materials; and 2) to test whether microwave data can be used for PSM, especially when other important co-variables are missing or insufficient.

6.2 Study site

The Waterloo aquifer is located within the Waterloo Moraine, about 10 km west of the city of Waterloo in southwestern Ontario (Figure 6-1). Annual average air temperature is 7ºC and mean annual precipitation is around 900 mm. The land use of the aquifer is mainly crop production, mixed with woodlots, feed lots and residential areas. The soils in the Waterloo region including the study site have developed on unconsolidated sediments derived from the action of continental glaciers several thousand years ago. The bedrock of limestone and dolomite in the area can directly influence soil development, especially in places where it is closer to the surface. During the great Pleistocene ice age, the southwestern peninsula of Ontario including the Waterloo region was subjected to glaciation from several directions in the Great Lake basins, so the glacial till in this region

is composed of unsorted or poorly sorted mixture of clay, silt, sand, gravel, and boulders (Presant and Wicklund, 1971). The diverse glacial till materials also affect the chemical and drainage characteristics of the soils, and the spatial pattern of glacial till in this region is directly linked to soil distribution. The Waterloo Moraine (also known as the Waterloo Sand-hills) consists mainly of sandy and silty deposits with occasional layers of clay and gravel, on moderately rolling topography. The surficial geology of the area is mapped at a scale of 1:50,000 (Ontario Geological Survey, 2010) and is used as part of the PSM training data (see details below).

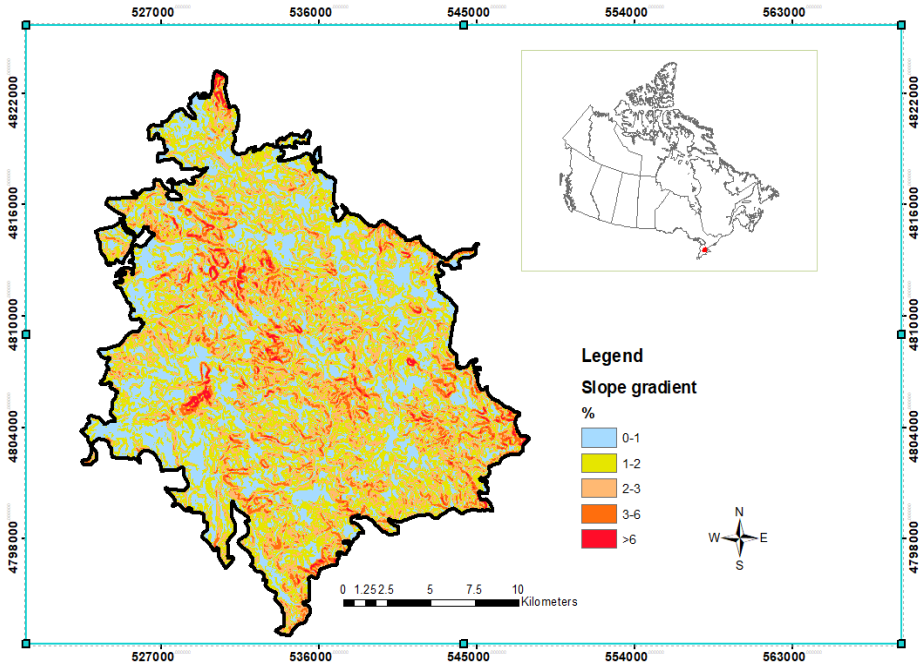


Figure 6-1 Waterloo aquifer study site, Ontario, Canada, showing slope gradient derived from Canadian Digital Elevation Model (Government of Canada, "Canadian digital elevation model 1941-2011." https://open.canada.ca/data/en/dataset/7f245e4d-76c2-4caa-951a-45d1d2051333, accessed May 7, 2020)

6.3 Data and methods

The following sections include data and data processing methods used in this chapter, however several data sets are from other cited sources. More detailed metadata and data lineage can be found in the original sources. Data processing was conducted using several open source tools such as SAGA (Conrad el al., 2015), Sentinel Application Platform (ESA, "SNAP." https://step.esa.int/main/toolboxes/snap/, accessed May 7, 2020) and R (R Core Team, 2018). RF was used to develop two soil maps of grouped soil series with common parent materials (e.g. surficial geology) to evaluate the utility of SAR data in PSM when surficial geology survey data is unavailable. One RF model included SAR data and the other included surficial geology data, while both models included surface characteristics (crop type, NDVI, DEM and its derivatives including MRRTF, MRVBF, plane and profile curvature, relative hill slope position, slope gradient, and topographic wetness index), as well as grouped soil series as described in the sections below.

6.3.1 Detailed legacy soil survey and training data mining

The latest detailed soil survey at 1:20,000 scale in Waterloo region was conducted during the soil survey glory period (between 1950 and 1995) (Anderson and Smith, 2011) and published in 1971 (Presant and Wicklund, 1971). In this detailed soil survey of the extended Waterloo region, 52 soil series were identified and mapped using orthorectified aerial photos, field sampling and laboratory analysis. The mapped soil polygons generally delineate 1 to 2 acres of land. Each polygon only contains one soil series. To report land-use suitability in the survey, topographic attribution was used to further annotate mapped soil series. For example, topographic slope gradient was defined with four classes: A (0-

3%), B(3-6%), C(6-12%) and D (>12%). Those slope classes were combined with the soil series name of the soil survey. This embedded knowledge was used to locate a sampling point within a mapped polygon. For example, a polygon annotated with HuB (Huron Loam with B class slope) provides more specific knowledge for locating the soil within that polygon. In the detailed soil survey in the Waterloo region, the 52 soil series reported at various drainage or topographic positions translate into about 200 combinations or map units, leading to many soil polygons with similar attributes (Figure 6-2). Those polygons with comparable map units were assigned similar soil physical and chemical properties. Although there were 52 soil series reported for Waterloo county, some of the soil series were either absent or rare.

Just like the legacy soil surveys elsewhere in Canada, soil series with similar parent materials are grouped and mapped with catena and association concepts. Within a catena unit, the member soils are distributed along the topographic and/or drainage gradients continuously. A catena is named according to the name of a naturally well-drained soil series of the catena. In this 1:20,000 detailed soil survey in Waterloo region (Figure 6-2 a), 20 soil catenas or associations were identified (Table 6-1). The soil series of the catenas were defined and mapped mainly through using tacit and qualitative knowledge. Soil series within a catena must develop on similar soil parent materials. For the objectives of this thesis chapter, the reported soil types were grouped and mapped into more generalized groups (Table 6-1; Figure 6-2b) based on reported parent material types and soil texture.

Based on the original detailed 1:20,000 soil survey (Figure 6-2a; Presant and Wicklund,

1971), two generalized soil maps were derived to study the differences caused by alternative soil series definitions. The first map preserves the one-to-many relationship between soil polygons with common names; it was produced by assigning all the polygons within the same map unit with a common catena identifier (Figure 6-2b). The second map was produced by grouping soil series based on catena names into generalized soil groups with the same parent material type (Table 6-2; Figure 6-2c). Polygons of a catena should have similar parent materials with different drainage conditions. So Figure 6-2c represents soil parent materials derived from the soil survey.

Table 6-1 Soil catena names and catena soil series, Waterloo county (Presant and Wicklund, 1971)

Soil Catena Name	Soil Series Name			
	Drainage			
	Good	Imperfect	Poor	Very Poor
Bennington	Bennington	Tavistock	Maplewood	
Bookton Bookton	Bookton Bookton	Berrien Wauseon	Wauseon	
Boomer	Boomer	Donald	Hawkesville	
Brant	Brant	Tuscola	Colwood	
Burford	Burford	Brisbane		
Caledon	Caledon	Camilla	Ayr	
Dumfries	Dumfries			
Farmington	Farmington	Brooke		
Fox	Fox	Brady	Granby	
Freeport	Freeport	Kossuth		
Grand	Grand	Macton	Elmira	
Guelph	Guelph	London		
Huron	Huron	Perth	Brookston	Dorking
Kirkland	Kirkland	Haysville	Hespeler	
Lisbon	Lisbon			
Mannheim	Mannheim			
St_Clements	St_Clements	Wellesley		
St_Jacobs	St_Jacobs	Floradale		
Waterloo	Waterloo	Heidelberg		
Woolwich	Woolwich	Conestogo	Maryhill	

This derived soil parent material map along with the independently collected 1:50,000 surficial geological material data (Ontario Geological Survey, 2010) were used to study the relationship between soil spatial patterns and time series remotely-sensed SAR images. To build the training point data for machine learning-based PSM with the co-variables as detailed in Methods 1 and 2 in Table 6-2, five training points per km^2 were randomly taken for each of the soil groups (Table 6-3). Soil groups representing very limited areas or that were altogether absent within the study extent did not have enough point training data, defined as those with less than 5 training data points (e.g. soil group 8), were dropped from further predictive mapping, leading to no pixels mapped for those classes. For the groups which could support more than 5 random sampling points, 100 random sets were generated to train the random forest algorithm. The random forest algorithm from the caret package in R was used (Kuhn et al., 2019). With 100 sets of randomly-mined training point data, 100 sets of soil group prediction results were produced. For each of the predicted pixel locations, among the 100 predictions, majority voting was applied to create the final classification map.

Table 6-2 Generalized soil groups using the 1:20,000 soil survey of Waterloo county

Group ID	Survey Reported Soil Names
1	Colwood loam, Elmira loam, Granby sandy loam, Hespeler sandy loam, Ayr sandy loam, Brookston loam, Brookston sandy loam Maplewood loam, Wilmot sandy, Wilmot silty clay loam, Wauseon sandy loam, Hawkesville loam
2	Perth loam, Perth sandy loam, Perth silty loam, Wellesley silty clay loam
3	Boomer loam, Guelph loam, St. Jacobs loam, Woolwich loam, Burford cobbly loam, Burford gravelly loam, Huron loam, Huron sandy loam, St. Clements sandy loam, Bennington loam, Brant loam, Caledon sandy loam, Freeport sandy loam, Huron clay loam, Maannheim loam, St. Clements silty clay loam, Guelph sandy loam, Kirkland sandy loam
4	Tuscola loam, Conestogo loam, Landon loam, Wellesley sandy loam, Brady sandy loam, Donald loam, Farmington sandy loam, Macton loam, Haysville sandy loam
5	Brisbane loam, Floradale loam, Kossuth sandy loam, Tavistock loam, Heidelberg fine sandy loam, Berrien sandy loam, Camilla sandy loam
6	Bookton sandy loam
7	Lisbon sandy loam, Fox sandy loam, Waterloo fine sandy loam
8	Grand loam, Martin sand and gravel
9	Organic soils

Table 6-3 List of co-variables used for Method 1 and Method 2

Method	List of co-variables
Method 1	SAR backscatter PCA components 1 and 2, crop type, NDVI, DEM, MRRTF, MRVBF, plane and profile curvature, relative hill slope position, slope, and topographic wetness index
Method 2	Surficial geological materials type, crop type, NDVI, DEM, MRRTF, MRVBF, plane and profile curvature, relative hill slope position, slope, and topographic wetness index

Table 6-4 Proportion of grouped soil types in the study area of Waterloo Aquifer

Soil group ID	Total area (km^2)	Count of samples
1	21.7	108
2	11.8	58
3	121.5	607
4	21.2	105
5	31.7	158
6	7.8	39
7	105.1	525
8	0.1	0
9	13.1	65

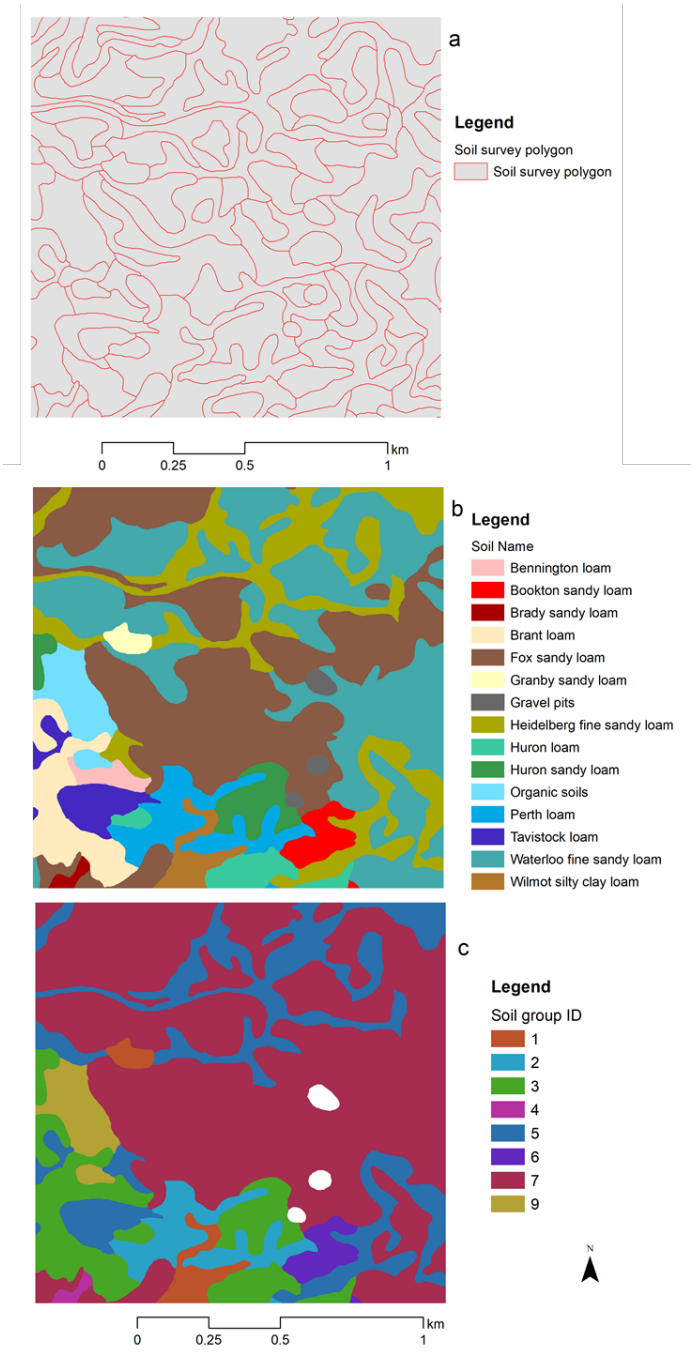


Figure 6-2. Sub-area of legacy and grouped soil type maps in the Waterloo County study site. a) legacy soil polygons, b) legacy 1:20,000 soil types (Presant and Wicklund, 1971); c) grouped soil type map with nine groups as described in Table 6-2. White patches are areas not mapped due to limited coverage of this soil group (see details in Section 6.3.1).

6.3.2 Paired synthetic aperture radar (SAR) data acquisition and processing

By referencing the publicly-available archived rainfall records in a study area, pairs of remotely-sensed SAR data can be found before and after rainfall events. SAR backscattering coefficients are affected by soil moisture content (soil dielectric constant), surface roughness, and vegetation cover, among other factors. However, before and after a rainfall event, the temporal variability of surface roughness and vegetation is lower than for soil moisture, and therefore, the change in SAR backscatter coefficients between repeat passes before and after a rain event result mainly from the changes in soil moisture (Moran, 2004). Is this remotely-sensed spatial pattern of soil moisture associated with soil types and underlying parent materials? In the context of this question, SAR data were used to study the utility of multi-temporal radar for PSM. For this study, Sentinel-1A Extra Wide Swath (EW) Level-1 Ground Range Detected (GRD) mode was used. The Sentinel-1A platform has C-band (3.8-7.5cm) wavelength, a twelve day revisit time, and the data produced are freely available from the Copernicus Open Access Hub (ESA, https://scihub.copernicus.eu/dhus/#/homeData, accessed May 7, 2020). To further avoid unwanted interference from crop/vegetation cover, the search time windows for the archived Sentinel-A SAR data were narrowed down to months before leaf out (e.g. May and June) or after leaf off and harvesting time (e.g. October and November). By referencing archived precipitation records before crop emergence (Figure 6-3), Sentinel-1A SAR data were retrieved as detailed in Table 6-5. Radiometric correction was conducted to calculate sigma naught from the intensity image. Further speckle smoothing using an Intensity Driven Adaptive Neighborhood (IDAN) filter and terrain correction were conducted. Details on SAR backscatter coefficients such as beta, sigma and gamma naught

calculations can be found in Small (2011). The final processed SAR data were resampled from 40 m to 30 m as part of the harmonization of the geospatial data for this study. Ratios of the wet and dry image pairs were calculated using Equation 6.1:

$$Diff - ratio = \frac{\sigma^0_{wet} - \sigma^0_{dry}}{\sigma^0_{dry}} \qquad \text{Equation 6.1}$$

where σ^0_{wet} and σ^0_{dry} are the sigma naught image of wet and dry conditions, respectively. Principal Component Analysis (PCA) was conducted among the four SAR images and the ratio images, to identify the main components which define the relationship between spatial patterns of soils and SAR signals. PCA analysis can also be used for feature reduction, especially when there are too many candidate input features for PSM. All the statistical procedures were conducted using R with Rgdal (Bivand et al., 2019), Raster (Hijmans, 2015) and SP (Pebesma and Bivand, 2005) libraries.

Normalized Radar Backscatter (Soil) Moisture Index (NBMI; Shoshany et al., 2000) was also derived using the wet and dry images (Equation 6.2):

$$NBMI = \frac{\sigma^0_{wet} - \sigma^0_{dry}}{\sigma^0_{wet} + \sigma^0_{dry}} \qquad \text{Equation 6.2}$$

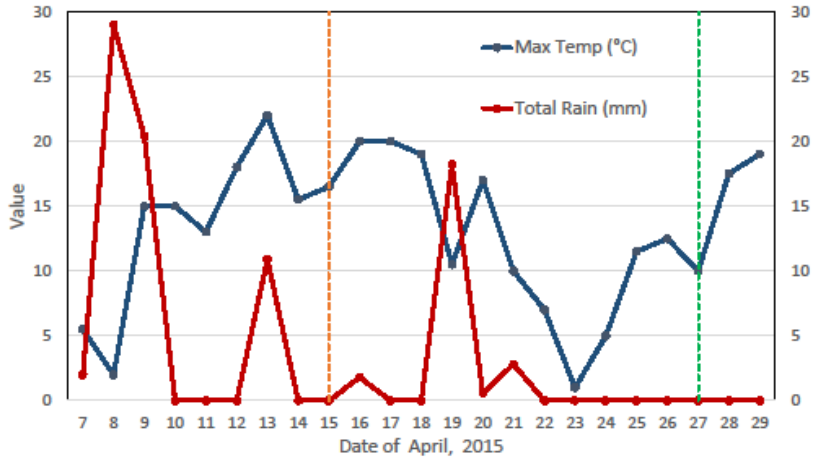


Figure 6-3 Study site temperature and total precipitation in April, 2015 (Government of Canada, "historic climate data. Last updated October 22, 2019." https://climate.weather.gc.ca/historical_data/search_historic_data_e.html, accessed May 7, 2020). The vertical lines intersect with the dates when SAR images were acquired.

Table 6-5 Details of Sentinel-1A SAR data used in this study

Mission	Product type	Acquisition date	Moisture status	Incidence angle	Pass
Sentinel-1A	GRD HH	2015-04-15	Wet (last rain on the 13th)	26.37 – 28.45	Ascending
	GRD HH	2015-04-27	Dry (last rain on the 21st)	26.38 – 28.46	Ascending

6.3.3 Other data sources used in this study

Landsat-8 imagery from May 2, 2015 was retrieved from the US Geological Survey web site (USGS, "Earth explore." https://earthexplorer.usgs.gov, accessed May 7, 2020). The Landsat-8 data were used to calculate Normalized Difference Vegetation Index (NDVI),

which was used to identify bare and sparsely vegetated patches for correlation analysis between microwave signatures and parent/surficial geological material types. In this study, bare or sparsely vegetated areas/patches were defined by NDVI <0.3.

Surficial geological material data were downloaded from the Ontario government web site (Ontario Geological Survey, 2010). The surficial geology of the Waterloo Aquifer was re-grouped to seven categories: Organic, Gravel, Sand, Fine to Very Fine Sand, Diamicton Sand, Diamicton Clay, and Silt. The numbers of the grouped surficial geology names are the numeric codes used in the GIS file. Within bare or partially bare areas, the microwave backscatter signatures were compared with those associated categories of the surficial geological materials.

The 2015 crop type map from the Canada Annual Crop Inventory published by Agriculture and Agri-Food Canada (Government of Canada, "Annual crop inventory, 2015." https://open.canada.ca/data/en/dataset/3688e7d9-7520-42bd-a3eb-8854b685fef3. accessed May 7, 2020) was used to generate a mask layer to exclude built-up and water-covered areas.

6.3.4 Spatial and multi-temporal covariates for PSM

To study the capacity of using paired SAR data for PSM, two sets of covariates were used in RF models. Set one included surficial geological material data without SAR data-derived covariates. Set two used paired SAR derived principal components without including surficial geological material data. Testing which set resulted in the more accurate soil map

was used to evaluate if SAR data could be used in PSM when surficial geological material data are lacking.

In addition, both RF models used the following covariates: generalized crop type data, DEM and its derivatives including MRRTF, MRVBF, plane and profile curvature, relative hill slope position, slope, and topographic wetness index.

6.3.5 Statistical analysis with unequal sample populations

In order to analyze the relationship between spatial patterns delineated by surficial geological material types and radar backscatter, the top two principal components (accounting for 86% of variance) of the time series radar data were used. The pixel values of the principal components data layer were sampled by each of the surficial geological material-defined polygons or patches. Within the study extent, all the "homogenous" surficial geological material patches were used for this analysis except for those covered by vegetation, leading to NDVI greater than 0.3. Higher vegetation cover leads to high noise in radar backscatter values. As a result, the numbers of available "homogenous" patches for each of the surficial geological material types were not equal; although the original SAR images were speckle filtered and smoothed, the SAR backscatter signature values may still have had outliers or extreme values. Therefore, within each of the sampled patches, the median of the principal component values was used for this analysis. Data distribution was evaluated with the Shapiro-Wilk test. If the median values were normally distributed, paired T-tests were used. Otherwise, nonparametric Kruskal-Wallis test (Hollander and Wolfe, 1973) with Dunn's post-test (Dunn, 1964) were performed. The

Kruskal-Wallis test, a non-parametric alternative to the one-way Analysis of Variance (ANOVA) test, was used to see if at least one surficial geology group was different from others based on the medians of the SAR principal component values. Following a significant Kruskal-Wallis test (if the null hypothesis of the Kruskal-Wallis test is rejected), the Dunn's post-hoc analysis test was performed to determine the SAR signature differences among the surficial geological material groups.

6.3.6 Validation data set

Forty-six points were inspected and sampled to verify the survey reported soil types and properties between 2013 and 2014. The validation points are not proportionally allocated to each of the grouped soils, so this validation data set had limitations, but some form of validation was valuable to help evaluate the methods.

Instead of using point to point validation, buffered neighborhood validation method was used. Specifically, pixels were considered to be correctly predicted if the validation point matched the pixel value within a pixel size radius (e. g. one pixel size of 30m).

6.4 Results

6.4.1 Backscatter of Sentinel-A SAR and surficial geological material types

The paired Sentinel-A microwave images as well as derived Normalized Backscatter (Soil) Moisture Index and wet/dry ratios were used to compute principal components and to study the relationship between the backscatter and predefined surficial geological material types. The first two principal components of the SAR features explain 86% of the spatial variance

of surficial geological material patterns (Table 6-6). The eigenvalue in Table 6-6 was used along with the eigenvector of the PCA output to calculate the component loading table (Table 6-7). From Table 6-7, the first PCA component mainly represents the contribution of the difference of the temporal SAR images, and the second one mainly corresponds to importance of the individual time series SAR images. The first two PCA components were used to analyze the spatial relationship between the microwave data and surficial geological or soil parent material data (Figure 6-4).

Table 6-6 Sentinel-1A PCA results showing eigenvalues and cumulative variance explained.

PCA outputs	Sentinel_PC1	Sentinel_PC2	Sentinel_PC3	Sentinel_PC4
Eigenvalue	2.04	1.41	0.52	0.03
Accumulative variation	0.51	0.86	0.99	1.00

Table 6-7 Principal Component Analysis (PCA) loading table. * Diff-ratio and NBMI were derived using Equation 6-1 and 6-2 respectively, with the two images acquired on April 15 and April 27, 2015.

PCA input data	Sentinel_PC1	Sentinel_PC2	Sentinel_PC3	Sentinel_PC4
Sentinel 1A (April 15, 2015)	0.26	0.83	0.50	0
Sentinel 1A (April 27, 2015)	-0.21	0.85	-0.49	0
Diff-ratio *	0.98	0.01	-0.13	0.13
NBMI *	0.99	-0.05	-0.10	-0.13

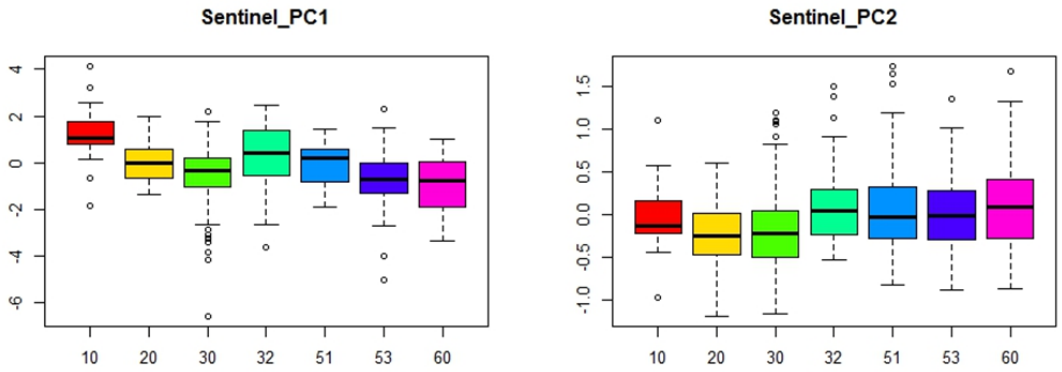


Figure 6-4 Boxplots of the PCA components and studied surficial geological material types. 10-Organic, 20-Gravel, 30-Sand, 32-Fine to Very Fine Sand, 51-Diamicton Sand, 53-Diamicton Clay, and 60-Silt.

The median values of the PCA components across surficial geological material patches were not normally distributed after the Shapiro-Wilk test. Due to the skewed sample distribution and unequal sample sizes of the PCA median values in association with the surficial geological material types, a Kruskal-Wallis test (Hollander & Wolfe, 1973) was used to investigate the statistical differences of the PCA median values in association with surficial geological material types. With the significance Kruskal-Wallis test results, the Dunn's test (Dunn, 1964), a post-hoc analysis, was performed to determine how the PCA median values by surficial geological material types were different from each other (Table 6-8). From Table 6-8, Sentinel-A backscatter data with 90% confidence level can differentiate the spatial patterns of surficial geological materials except for those closely related geological material types. For example, there was confusion between category 32 (Sand to Very Fine Sand) and 51 (Diamicton Sand). Another example of confusion is between 53 (Diamicton Clay) and 60 (Silt). Again, the unsorted nature of Diamicton clay may have caused confusion between the two material types. Merging some of those closely

related material types is recommended.

Table 6-8 Dunn's test between surficial geological materials and Sentinel-A SAR PCA1(alpha = 0.1). T(True) indicates significant difference; F(False) means no significant difference.

Surficial geological material type	20 Gravel	30 Sand	32 Sand-Very Fine Sand	51 Diamicton Sand	53 Diamicton Clay	60 Silt
10 Organic	**T**	**T**	**T**	**T**	**T**	**T**
20 Gravel		**T**	F	F	**T**	**T**
30 Sand			**T**	**T**	**T**	**T**
32 Sand-Very Fine Sand				**T**	**T**	**T**
51 Diamicton Sand					**T**	**T**
53 Diamicton Clay						F

6.4.2 PSM with SAR-derived covariate and surficial geological material data

To further demonstrate the potential use of SAR derived covariates for PSM, the Random Forest algorithm was used to predict soil types based on two sets of identical input co-variables except for the mutually exclusive use of surficial geological material and SAR derived covariate data. Figure 6-5 shows the predicted soil groups with surficial geological material data on the left (a) and with SAR-derived covariates on the right (b). Table 6-9 shows the summarized distribution of the internal accuracies across the 100 machine learning runs. The internal accuracy values in percentage in Table 6-9 were calculated from 30% of the total samples to cross-validate the RF models. Unlike the out of bag sample values that are used for internal cross-validation during the RF model building stage, these withheld samples provide an independent and more robust assessment of RF model performance (Kuhn, M., 2019).

The values from Table 6-9 are mainly used to understand the stability and consistency of the 100 machine learning runs. With the RF models built from the training data, there is 64% agreement between the predicted soil maps from the two sets of covariates when per-pixel comparison was conducted. This agreement is further detailed in Table 6-8, which summarizes how each of the predicted soil groups agreed and disagreed between the two predictions; the major confusions were between groups 3 and 7, and 1 and 5, all of which contain soils with loamy textures (see Table 6-3). However, the true validation of the prediction accuracies needs to be based upon independent validation point data. In this study, 46 independently sampled point data were used to validate the predicted soil group maps. The overall accuracy of the predicted soil groups using surficial geological material data and SAR-derived data was 62% and 64% respectively. Therefore, when surficial geological material data are too coarse in resolution, or lacking in availability, high resolution SAR-derived covariates can be used in PSM.

Table 6-9 Summarized internal accuracies of 100 RF runs and two RF models using surficial geology (SG) vs. SAR data

	Using SG data %	Using SAR data %
Minimum	42	38
1st quantile	45	40
Mean	45	41
3rd quantile	46	41
Max	49	43

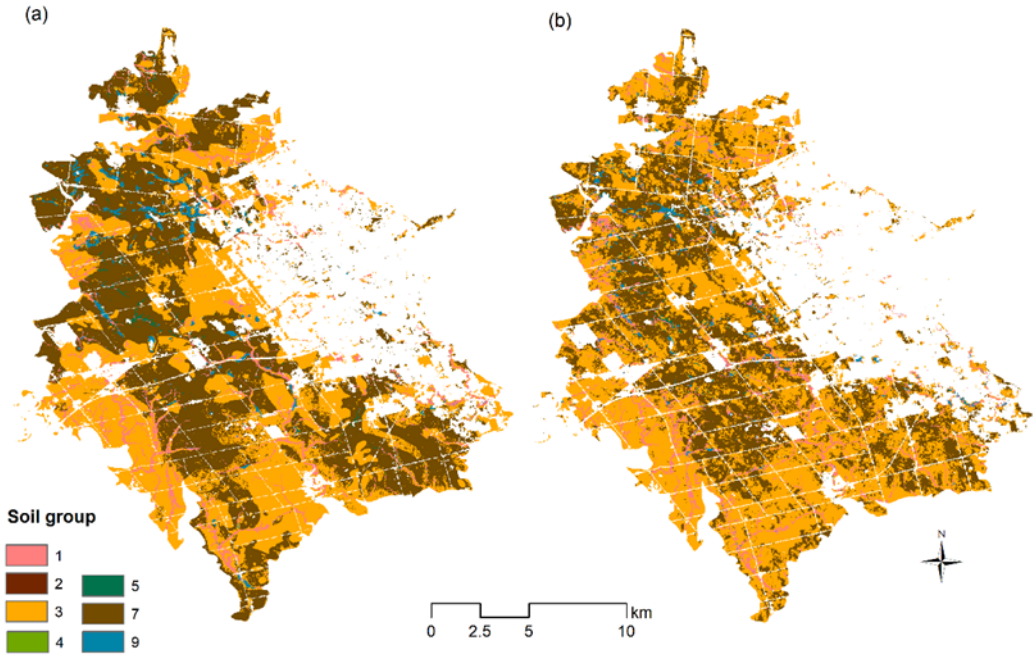


Figure 6-5 Predicted soil groups in Waterloo Aquifer: (a) predicted soil groups using surficial geological material data; (b) predicted soil groups using SAR derived covariates. Detailed soil group information is found in Table 6-3.

Table 6-10 Per-pixel comparison between the predicted soil groups. The values in the table are the pixel counts within the mapped extent.

Soil group	1	2	3	4	5	6	7	8	9
1	13747	0	1656	0	127	0	391	0	984
2	0	0	1	0	0	0	0	0	0
3	1224	0	96833	1	151	0	29014	0	216
4	24	0	2	0	0	0	0	0	0
5	954	0	179	0	447	0	180	0	26
6	0	0	0	0	0	0	0	0	0
7	1848	0	61441	0	299	0	70674	0	615
9	1070	0	1071	0	28	0	1067	0	2123

6.5 Discussion

In Canada, surficial geological materials greatly influence soil development and distribution (Geng et al., 2010b). In this chapter, the key objective was to study if remotely-sensed microwave or SAR data can be used to delineate spatial patterns of soil parent material. The use of paired MODIS remotely-sensed data for PSM has been studied in Manitoba, Canada (Liu et al., 2012) as well as in northeast China (Zhu et al., 2010a). With similar methodology used by Liu et al. (2012), it was hypothesized that the response surfaces of remotely sensed SAR data over wet and dry cycles correlated with the spatial pattern of soil parent materials. This hypothesis was shown to be correct for the Waterloo aquifer region in Ontario with 64% agreement between pixels of the predicted soil group maps composed of 6 different soil parent materials.

The results of this study corroborate those of wetland/peatland studies, which also showed the potential for SAR data to delineate (peatland) soil spatial distributions in Canada (Millard and Richardson, 2013; Li and Chen, 2005) The soils investigated in this study were primarily mineral soils in a temperate climate. To apply these methods to primarily organic soils or permafrost soils, further research is needed. For example, organic soils, especially peatland soils, are distributed along biogenic landscapes (Kroetch et al., 2011). The biogenic landscapes can be further detailed with DEM data. Regardless of organic or mineral soil mapping objectives, the proper choice of DEM resolution in combination with SAR data needs further study, as finer resolution is not always better (Minasny et al., 2019). For example, different resolutions and sources of DEMs did not increase the peatland mapping accuracy in Minnesota (Knight et al., 2013). For PSM in places where landforms are more homogenous, coarser resolution elevation data can be sufficient as one of the co-variables to effectively predict soil types and properties (Cavazzi et al., 2013).

In this study, the paired SAR data were collected from leaf-off periods. However, surface roughness due to vegetation cover cannot be avoided completely, particularly for agricultural regions where there are few periods with fallow conditions and for non-agricultural regions where there is permanent vegetation cover. Longer wavelength microwave data (e.g. L and P band SAR) have the ability to penetrate vegetation cover better and as a result, are expected to be less noisy (Karthikeyan et al., 2017). There is need to further explore the utility of long wave length microwave remotely-sensed data for PSM, which is now becoming available with the launch of ALOS-4 (JAXA, “Advanced Land

Observing Satellite-4 (ALOS-4)." https://global.jaxa.jp/projects/sat/alos4/, accessed May 7, 2020). With more accessible sensors and multi-polarization SAR data in Canada, the effectiveness of different polarization microwave and other kinds of remotely sensed data also need to be evaluated for operational use in PSM applications (Minasny et al., 2019).

Another issue is the need for shorter satellite revisit times (e.g. daily) to better synchronize the SAR data acquisition dates with the wet-dry cycle of weather events. This is needed because single satellite platforms such as RadarSat-2 often have longer satellite revisiting time (e.g. maximum 24 days). The Radar Constellation Mission (RCM) will provide shorter time interval (e.g. less than 4 days) for repeating satellite scenes for a study area (Canadian Space Agency, "RADARSAT Constellation Mission." https://www.asc-csa.gc.ca/eng/satellites/radarsat/default.asp, accessed May 7, 2020).

As used in this chapter, PCA methods can be used for feature reduction especially when many time series satellite images are used. There are other feature reduction and selection methods, such as feature importance ranking with RF (Hengl et al., 2017), the Best General Linear Model (McLeod and Xu, 2018), and multi-scale feature analysis (Behrens et al., 2018a, 2019). Feature reduction is important for predictive soil mapping especially in terms of effectively managing computing resources. For example, for nationwide PSM at 250 m resolution in Canada, available input features such as remotely-sensed, climatic, and DEM derived data can have very high volumes; on average, each input feature can require 2 GB. To avoid associated computation impediments, it is important to statistically reduce the number of input features for PSM. As PCA is just one of the effective options for feature

reduction, the pros and cons of different feature reduction measures for PSM should be further studied.

Although soil parent material data are part of the national soil database, these data are presented at a variety of scales and resolutions (Schut et al., 2011). In places where detailed soil surveys exist (such as 1:20,000 or larger), soil survey-derived soil parent material map data can be used for predictive soil mapping. However, such detailed soil surveys are very limited in Canada. Coarser soil survey polygons often include many soils or components and cannot be used to derive spatially explicit soil parent material map data. Soil parent material map data derived from those coarser soil polygons is not reliable, especially for finer resolution (e.g. < 30 m grid size) PSMs. Can we use various scales of surficial geological material map data to derive soil parent material map data? The answer is yes, but with some challenges. The first challenge is that the legends or classification schemes used in surficial geological materials map data are defined by the Geological Survey of Canada (Government of Canada, "Surficial Geology: Open government portal" https://open.canada.ca/data/en/dataset/cebc283f-bae1-4eae-a91f-a26480cd4e4a, accessed May 7, 2020) and are different from those used to name soil parent materials across the regions in Canada defined by the Canadian System of Soil Classification (Canada Department of Agriculture, 1973). Other challenges include highly variable scales of data across Canada, a broad range of vintages, and general lack of harmonization across the provinces and territories. For mapping areas larger than the extent of one satellite image scene, image normalization of multiple scenes will be a new challenge. When and where finer resolution surficial geological material data are lacking, purposively selected time

series SAR derived data may be used for PSM as demonstrated in this chapter. With shortened satellite revisiting time from the Radar Constellation Mission (RCM), acquiring SAR data immediately before and after major rainfall events will become practical. It is expected that the application of multi polarization and longer wavelength SAR data for PSM will also be broadly extended.

6.6 Conclusions

Surficial geological or soil parent materials often dictate soil distribution and spatial patterns of the soils in Canada. The results of this study suggest that when and where finer resolution surficial geological or soil parent material data are lacking, purposively selected time SAR images and derived data can be used for PSM. However more study is required especially on using longer wavelength and multi-polarization microwave remotely-sensed data. With shortened satellite revisit times from the Radar Constellation Mission (RCM) in Canada, acquiring SAR data immediately before and after major rain fall events will become more practical. The research reported from this chapter will form the basis for larger areas including nationwide PSM and soil landscape resource inventory and accounting. However, more research is needed to extend these methods to regions with permafrost and for organic soil carbon accounting related issues.

Chapter 7: **Synthesis: A framework for predictive soil mapping in Canada, conclusions and future research needs**

7.1 Introduction

Conventional soil surveys have been conducted across Canada for decades and have provided valuable information for land management, environmental assessment, ecosystem modeling and other uses. However, the investment to continue conventional soil surveys has been drastically decreased from various levels of government and private sectors. With recent advances in geospatial data science, high performance computing and inference methods, cost-effective and reproducible soil and soil landscape mapping is possible via a PSM framework. From the research described in previous chapters, a recommended PSM framework for operational use has been developed, including recommended data and knowledge mining, purposive sampling design, field sampling and measurements, laboratory analysis, sample and soil covariate data preparation, training of selected machine learners, ensemble inference, validation and output data management and use. This chapter presents more details of several key components of the recommended PSM framework, summarized in Figure 7.1. The framework was developed based on a combination of lessons from previous studies, as reviewed in Chapter 2, and the findings of Chapters 4-6. This framework is now being used operationally in Canada.

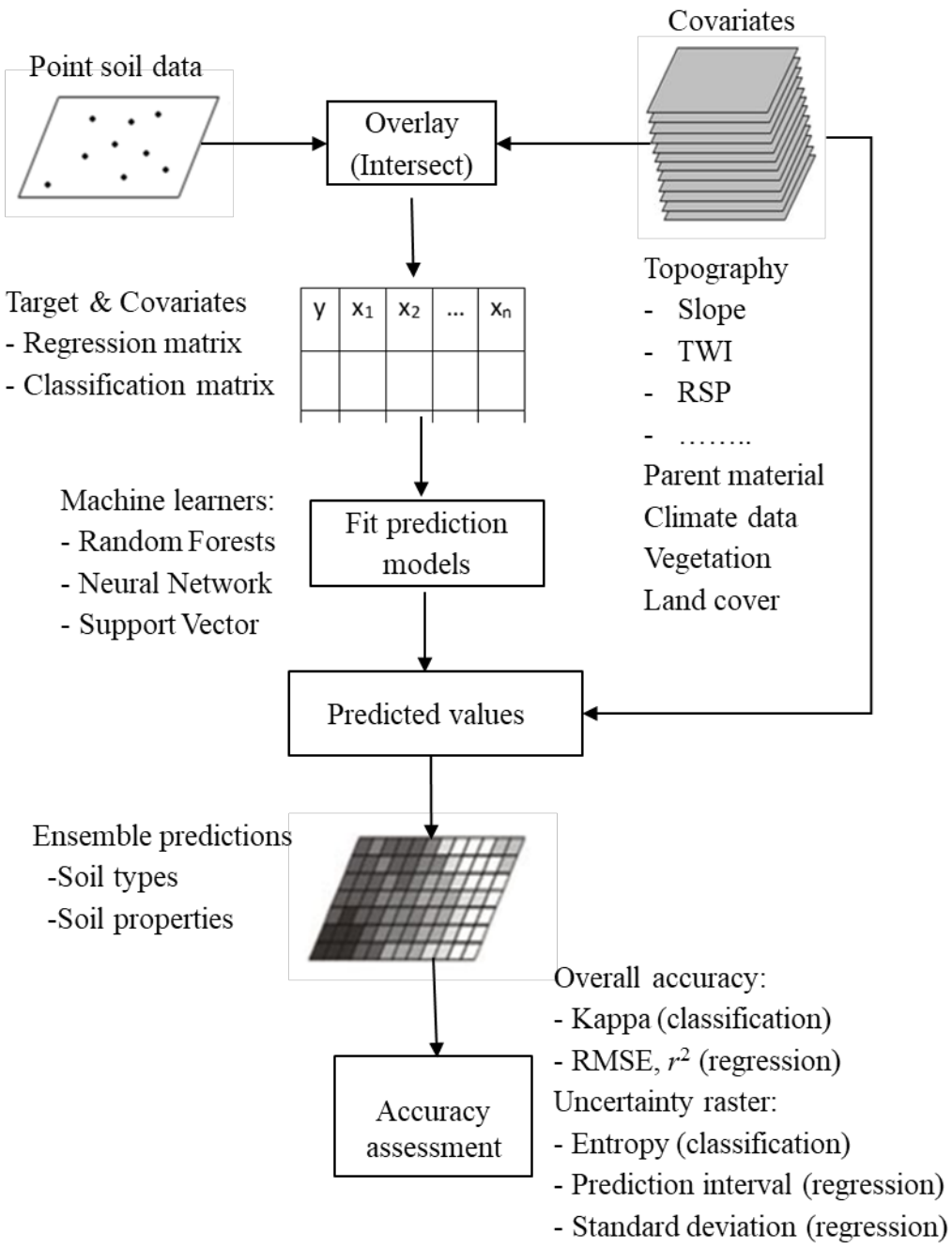


Figure 7-1 - Framework for machine learning based predictive soil mapping for Canada (modified from Shangguan et al., 2017)

7.2 Operational predictive soil mapping in Canada

In Canada, all conventional soil surveys have been completed with the goal of supporting practical uses including agricultural land use (Schut et al., 2011). This has implications for mapping scale, field work planning, and resource budgeting. Predictive soil mapping should follow similar steps by first identifying project requirements. It is important to research and develop predictive soil mapping methods contextually. For example, to map soil properties such as texture and soil carbon at watershed and national scales, the spatial resolution of input and output data using the PSM framework can be finer for watershed scales, and coarser for national extents. Finer resolution data are often more expensive to produce than coarser versions, and excessively fine resolution data produces unwanted noise. However, coarser resolution input data may not be able to produce the expected detail. Through this and other studies (Behrens et al., 2018a, 2018b), it is suggested that co-variables used as inputs for a PSM framework be selected and used across resolutions or scales. Using multiple themes and resolutions, more input variables may be available than required by the machine learning algorithms. Efforts can be made to reduce the number of input features through importance ranking statistics. Machine learning approaches like RF can rank the importance of input features; all are still used by default, but further analysis could choose subsets based on the ranking. Other options such as the Best General Linear Model (McLeod and Xu, 2018) can serve the need to reduce the number of input features as well. Different resolutions of PSM will also require different numbers of point training samples. The finer the resolution of output data, the more point sample data will be required for training machine learner approaches (further discussed in Section 7.6).

7.3 From geostatistics to ensemble machine learning

When there are enough statistically sampled point data, ordinary geostatistics such as kriging, linear regression, and even inverse distance weighting can be used to produce continuous area class or soil property maps. For larger mapping areas with less available sample data, machine learning-based approaches are recommended. Each of the frequently used machine learning methods such as RF, ANN, SVM and C5 performs with advantages and disadvantages in PSM, therefore ensemble averaging and optimization is recommended (Hengl and MacMillan, 2019). From Chapter 4 and 5, in order to study methods of PSM for both soil classes and soil properties, the key requirements for ensemble prediction are to prepare a common input dataset including both training and covariate data. The covariate data should be harmonized to the same extent, projection, and grid size. Categorical (e.g. parent material map) and continuous (e.g. slope gradient) input data types should be treated with care since not all the machine learners can accommodate both of the data types equally. Not all the machine learners will produce probability or uncertainty information of the predictions. The need to have uncertainty information presented along with the predicted soil maps should be dealt with by focusing on those methods which can produce uncertainty information. Uncertainties could also be captured during the validation stage using validation residual ranges and spatial distribution. As illustrated in Chapter 4, to produce final prediction maps using the ensemble approach, majority voting for categorical data and averaging for continuous data types can be used. However there are other optimization algorithms that should be further studied such as optimization according to distance to access and optimization by avoid build up areas. It is expected that ensemble

algorithms can improve on a simple averaging approach and therefore improve the prediction accuracies (Hengl and MacMillan, 2019).

7.4 Prediction of soil classes and properties using data mining and new data collection

Machine learning for PSM can incorporate soil expert knowledge as training data. Some of this expert knowledge is contained in legacy soil survey maps, reports and databases. Chapter 3 illustrates how existing soil surveys report soil classes and the associations between soil and environment. This legacy information can be used to understand and to define soil classes to be mapped using PSM. Quite often soils with different series names may possess similar soil genesis pathways. In Chapter 4, I illustrated the steps to understand and regroup those reported soil series for PSM. The grouping principles should follow the definitions of soil association and catena used in Canada. Within a map area, soil with similar parent materials should be grouped to associations. Within each association, soils should be grouped and differentiated by topographic sequence which is associated with drainage conditions.

In locations where very detailed soil surveys exist, those single component soil polygons could be used to locate the reported soil type spatially, which in theory can be used to train machine learners (Chapter 4). However data mining methods such as random vs stratified random sampling within a single soil type polygon still need further study. For example, a point being located in a single component soil survey polygon does not mean that it fully carries the central characteristics of the mapped soil. A random sampling method within a

single soil component polygon worked well in terms of soil type mapping in Chapter 4, however that approach may be challenging in some contexts, such as legacy soil surveys with coarser scales.

In countries like Canada where the soil classification system and soil survey are based on soil biogeographic genesis theory, the relationship between named soil types and environmental covariates is more direct than that between the properties and the covariate. So the preferred method to predict both soil class and properties using PSM methods is to map targeted soil types first and then to indirectly map soil properties after. The best use of legacy soil survey information is to use the pedon data as training or validation data sets for machine learning algorithms. To gather and harmonize as many points as possible is an ongoing effort in PSM. Regardless, the use of the legacy soil pedon data has several constraints. First, many of the surveyed soil pedons have not been statistically designed; the existing distribution of the pedon data tends to be skewed and clustered. Secondly, soil samples measured across many years and organizations have not been completed with the same or correlated laboratory procedures. A pH 6.5 for an A horizon sample is not necessarily different from pH 6.9 reading from another A horizon soil sample if the pH measurements came from different laboratory procedures. Finally, but not exhaustively, soil is a dynamic natural body. It changes over time. Some soil properties measured thirty years ago will not be the same today. So inferences based on trained machine learners using legacy point data will not be accurate in many cases. That means new point soil samples should be planned and taken as part of PSM work. Questions of where and how many new soil samples need to be taken for higher quality PSM have been dealt with in this study

using an optimized conditioned Latin Hypercube Sampling approach (Chapter 4 and 5). In practice, due to land access and other constraints, not all planned sample points can be taken. Also, repeating a sampling plan over the designed sampling locations requires accurate GPS readings. For added soil property measurements such as total phosphorus, soil sample DNA sequencing, or soil spectral scanning using near infrared spectrometry, new soil sampling protocols need to be developed.

7.5 New data sources and feature reduction

When co-variables closely related to soil development such as soil parent material are readily available for PSM, other geospatial data sources may not be needed. However, when the needed covariate data are lacking, other geospatial data, especially various sources of remotely-sensed data, can become viable alternatives. In Canada, the 1:7,000,000 surficial geological material map is much too coarse for finer resolution PSM work. In regions where topography is level or near plain, frequently used topographic derivatives also don't contribute much to the PSM. Remotely-sensed optical and microwave data collected before and after a major rainfall event have spatial patterns which are related to the spatial patterns of soil classes and soil properties. In Chapter 6, derivatives of wet/dry paired remotely-sensed SAR imagery showed potential due to imaging capability under all weather conditions. With more and more freely available earth observation data, the use of remotely-sensed data for PSM will increase dramatically in the future. Also dependent on the extent and resolution of a PSM project, some co-variables may be more important in one case rather than in others. For example, climate information, such as monthly mean gridded temperature data, is important for Canada-wide soil

mapping but not for a watershed PSM; information of topographic sequences derived from DEM is more important for a finer resolution (< 30m grid) PSM than at a watershed extent.

The total amount of environmental co-variables can grow quickly. For example, a DEM can be used to derive multiple landscape facets at various resolutions or scales. With added optical, near infrared and infrared, microwave and other spectral range remotely-sensed data at various resolutions, readily available co-variables for a PSM project need to be ranked and reduced using a feature reduction approach. Feature reduction will also improve computing capacity and speed. In Chapter 6, feature ranking using RF and feature reduction using PCA are used. Other feature reduction measures either from machine learning algorithms or from standalone methods such as the Best General Linear Model (BGLM) do exist.

7.6 Prediction accuracy, validation and uncertainty

The first step of accuracy assessment is to check the difference between observed and predicted values. This difference is also called residual analysis. For example, in Chapter 5, validation of PSM of BD, an important soil property, was carried out using the spline interpolated BD values at 0-10 cm, derived from the field sampled BD. Before using the residual values for further analysis, normality assessment of the overall residual values is needed. The normality check of the residuals can be conducted using Shapiro-Wilk Normality test. Skewed distributions of residual data suggest that the residuals need to be transformed and retested. Point to point validation is often conducted based on the goodness of fit between the observed and predicted BD values.

Predictive soil mapping with machine learning often includes two types of validation methods: model- vs. independent sample-based (Brus et al., 2011). The model-based validation is often done as part of machine learning procedures which include using split or leave one out cross validation (LOOCV) methods (Hengl and MacMillan, 2019). Although the assessment of the outcomes of PSM should not solely rely on the model-based validation, low accuracy of a prediction based on modeled validation will not lead to a better prediction. In Chapters 4, 5 and 6, both model- and independent sample-based validation were used. Questions on how many validation sampling points are required and where they should be located need to be studied further. For a small predictive soil mapping area, intensive validation points could be sampled as was done in Chapter 4 and 5 for soil survey reported soil type validation and soil properties such as BD measurements. To independently validate predicted soil maps for regional and national extents, the numbers of the required samples can be too large to sample feasibly. Alternative validation data sources such as using remote sensing data-derived pseudo points have been proposed (Lagacherie et al., 2019).

The purpose of conducting PSM is to develop the needed soil data for integrated use and decision making such as ecosystem modeling, sustainability metrics, and ecosystem goods and service evaluation. Uncertainty measures with prediction and validation procedures are important for understanding and quantifying the overall accuracy and reliability of integrated models and applications. Uncertainties from PSM should propagate into downstream calculations using pedo-transfer functions and model algorithms (Dobarco et

al., 2019). For example, uncertainty in the form of probability of soil class distribution is used to indirectly predict soil BD in this study. In the cases when PSM has high uncertainties, improving the prediction accuracy must become a very specific research objective.

7.7 Towards operational PSM in Canada

After reviewing the main PSM methods that have been studied and used across field, watershed, regional, national, and global scales, predictive–based machine learning methods were further studied. Machine learning methods all need field soil point data and/or expert knowledge. This information can be extracted and mined using legacy soil survey databases. However due to the limited availability, biased distribution and out-of-date of legacy point soil data, operational PSM should strive to collect new point soil data using appropriate statistical designs. Conditioned Latin Hypercube sampling design (as described in Chapter 5) is the recommended sampling design method, and it has been used in some real cases in Canada already. Some ecosystem models and global soil partnerships (Arrouays et al., 2017) also require vertical soil properties with uniform depth. For uniform depth soil property inference, the values of point soil data from sampled layers or horizons were modeled and interpolated using a Spline function (Bishop et al., 1999; https://github.com/cran/GSIF/blob/master/R/mpspline.R accessed May 7, 2020). Chapter 5 reported a method to derive soil properties via predicted soil types along with prediction uncertainties. The predicted soil property values were validated using independent validation point data. As described in Chapter 5, the residual values produced from validation step needs to be interpolated and examined if there is quantifiable and auto-

correlated portion of the residuals to be included for the final predicted soil maps.

Canadian landscapes are young and have evolved under the influence of the latest glaciations (Presant and Wicklund, 1971). The surficial, especially the glaciated surficial, geological materials dictate soil development and distribution across Canadian landscapes. To conduct PSM in Canada, surficial geological materials which constitute soil parent materials should be considered. However when the surficial geological material data are missing or lacking, alternative data sources, such as wet/dry paired remotely-sensed SAR images, are recommended. In Chapter 6, data derived from paired wet/dry SAR images were successfully correlated to surficial geological materials, and produced RF-based soil class maps with similar accuracy to PSM with surficial geological materials data. Further studies are suggested in the context of longer wavelength SAR data and recently launched Canadian radar constellation satellites. Longer wavelength SAR data will be less noisy than shorter wavelength SAR image in this regard because there will be less interaction with surface roughness and vegetation cover.

The primary goal of this thesis is to research and develop components of a framework for operation PSM in Canada. The framework developed in this thesis reflects similar efforts in Australia (Bui, 2006), Africa (Hengl et al., 2015) and European Union nations (Carré et al., 2007). The advantage of this framework is that Canadian training data and co-variables are treated in the context of Canadian landscapes and Canadian soil classification system. While more research is still needed, particularly in terms of environmental variable selection, optimized sampling design for new data collection, uncertainty measures and

validation methods, the methods and procedures of PSM from this thesis work are now being used in Canada for PSM from national to regional and watershed scales. Nationwide PSM efforts are contributing to Canadian's international agreements such as the United Nations global soil partnership initiative (FAO, "Global soil partnership." http://www.fao.org/global-soil-partnership/en/, accessed May 7, 2020), and regional and watershed scale predictive soil mapping methods are used across Canada for beneficial management practices (BMPs) and AAFC's Living Laboratory program (Government of Canada, "Living laboratory initiative: Agriculture and Agri-Food Canada." https://www5.agr.gc.ca/eng/science-and-innovation/living-laboratories-initiative/?id=1551383721157, accessed May 7, 2020). The optimized conditioned Latin Hypercube sampling method is being widely used and further studied across many watersheds from Ontario, Quebec and to PEI. With the knowledge and expertise acquired from this study, immediate application of this soil data development framework will be applied to national predictive soil mapping at finer (100m) resolution. Therefore, this research has already contributed many benefits to soil mapping in Canada.

References

Adhikari, K., Kheir, R.B., Greve, M.B., Bocher, 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 Science Society of America Journal 77: 860-876.

Adhikari, K., Minasny, B., Greve, M. B., Greve, M. H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma 214, 101–113.

Arrouays, D., Lagacherie, P., Hartemink, A, E, 2017. Soil mapping across the globe. Geoderma http://dx.doi.org/10.1016/j.geodrs.2017.03.002

Arrouays, D., McBratney, A.B., Minasny, B., Hempel, J.W., Heuvelink, G.B.M.,MacMillan, R.A., Hartemink, A.E., Lagacherie, P., McKenzie, N.J., 2014. The GlobalSoilMap project specifications. In: Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, Taylor and Francis Group, London, pp. 494.

Anderson, D.W., and Smith, C.A.S., 2011. A history of soil classification and soil survey in Canada: Personal perspectives. Canadian Journal of Soil Science, 91(5), 675-694.

Akumu, C.E.; Johnson, J.A.; Etheridge, D.; Uhlig, P.; Woods, M.; Pitt, D.G.; McMurray, S. 2015. GIS-fuzzy logic based approach in modeling soil texture: using parts of the Clay Belt and Hornepayne region in Ontario Canada as a case study. Geoderma 239: 13-24.

Arnold, J.G., Bieger, K., White, M. J., Srinivasan, R., Dunbar, J. A., and Allen, P. M., 2018. Use of decision tables to simulate management in SWAT+. Water. 10(6). Doi: 10.3390/w10060713.

Baghdadi, N., Cerdan, O., Zribi, M., Auzet, V., Darboux, F., Hajj, M. E., and Kheir, R. B., 2008. Operational performance of current synthetic aperture radar sensors in mapping soil surface characteristics in agricultural environments: application to hydrological and erosion modelling. Hydrol. Process 22:9–20

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

Beguin, J., Fuglstad, G., Mansuy, N., Paré, D. 2017. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma 306 (2017) 195–205

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmidt, M., 2005. Digital soil mapping using artificial neural networks. Journal of Plant Nutrition and Soil Science 168, 21–33.

Behrens, T., MacMillan, R. A., Viscarra Rossel, R. A., Schmid, K., Lee, J. 2019. Teleconnections in spatial modelling. Geoderma 354:113854

Behrens, T., Schmid, K., MacMillan, R. A., Viscarra Rossel, R. A., 2018a. Mulitscale contextual spatial modeling with Gaussian scale space. Geoderma 310:128-137

Behrens, T., Schmid, K., Viscarra Rossel, R. A. , Gries, P., Scholten, T., MacMillan, R. A., 2018b. Spatial modelling with Euclidean distance fields and machine learning. European Journal of Soil Science, doi: 10.1111/ejss.12687

Behrens, T., Zhu, A-X, Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155:175-185

Benjannet, R., Khiari, L., Nyiraneza, J., Thompson, B., He, J., Geng, X., Stiles, K., Jiang, Y., Fillmore, S., 2018. Identifying environmental phosphorus risk classes at the scale of Prince Edward Island, Canada. Can. J. Soil Sci. 98: 317–329, dx.doi.org/10.1139/cjss-2017-0076

Bezdek, J. C., Ehrlich, R., Full, W., 1984. FCM:the fuzzy c-mean clustering algorithm. Computers and Geosciences 10:191-203

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45.

Biswas A, Zhang Y. 2018. Sampling designs for validating digital soil maps: A review. Pedosphere. 28(1): 1–15.

Boruvka, L., Penizek, V., 2006. Chapter 30 A test of an artificial neural network allocation procedure using the Czech soil survey of agricultural land data. Developments in Soil Science 31:415-424

Breiman, L., 2001. Random Forest. Machine Learning 45:5-32

Breiman, L., Friedman, J. H. Olhsen, R. A., Stone, C. J., 1984. Classification and Regression Trees. Wadsworth, Monterrey, CA

Brevik, E. C., Calzolari, C., Miller, B. A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A., 2016. Soil mapping, classification, and pedologic modeling: History and future directions, Geoderma (264):256-274

Brus, D. J., Bogaert, P., Heuvelink, G, B. M., 2008. Bayesian maximum entropy prediction of soil categories using a traditional soil map as soft information. European Journal of Soil Science 59:166-177

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62, 394–407

Bui, E. N., 2004. Soil survey as a knowledge system. Geoderma 120 (1–2), 17–26.

Bui, E., 2006. Chapter 2 A Review of Digital Soil Mapping in Australia.In Digital soil mapping: An introductory perspective. (ed) Lagacherie et al., Elsevier, New York, 31:25-37

Bui, E.N., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecol. Model. 191, 431–446

Bui, E.N., Henderson, B.L., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Glob. Biogeochem. Cycle 23, GB4033. http://dx.doi.org/10.1029/2009GB003506.

Bui, E.N., Loughhead, A., Corner, R., 1999. Extracting soil–landscape rules from previous soil surveys. Australian Journal of Soil Research 37, 495–508.

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modeling and legacy data. Geoderma 103, 79–94.

Bulmer, C.E., Schmidt, M.G., Heung, B., Scarpone, C., Zhang, J., Filatow, D., Finvers, M.,Smith, C.A.S., 2016. Improved soil mapping in British Columbia, Canada, with legacy soil data and Random Forest. Digital Soil Mapping Across Paradigms, Scales and Boundaries. Springer Environmental Science and Engineering, pp. 291–303.

Burrough, P.A. 1989. Fuzzy mathematical methods for soil survey and land evaluation. J. Soil Sci. 40:477–492.

Burrough, P.A., 1992a. The development of intelligent geographical information systems. Int. J. Geogr. Inf. Systems 6, 1-12.

Burrough, P.A., MacMillian, R.A., van Deusen, W., 1992b. Fuzzy classification methods for determining land suitability from soil profile observations and topography. J. Soil Sci. 43, 193–210.

Burrough, P. A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77:115-135.

Camera, C., Zomeni, Z., Noller, J. S., Zissimos, A. M., Christoforou, I. C., Bruggeman, A., 2017. A high resolution map of soil types and physical properties for Cyprus: Adigital soil mapping optimization, Geoderma (285):35-49

Canada Department of Agriculture. 1973. Revised system of soil classification for Canada. Queen's Printer for Canada, Ottawa, ON.

Carré, F., McBratney, A. B., Mayr, T., Montanarella, L., 2007. Digital Soil Assessments: beyond DSM. Geoderma, 142:69-79

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma 195, 111–121.

Chang, D. and Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural network. Remote Sens. Environ. 74:534-544

Cianfrani C., Buri A., Verrecchia E., Guisan A., 2018. Generalizing soil properties in geographic space: Approaches used and ways forward. PLoS ONE 13(12):e0208823.https://doi.org/10.1371/journal.pone.0208823

Coen, J. M., 1987. (Ed) Soil Survey Handbook. Vol. 1 Technical bulletin 1987-9E, ISBN 0-Agriculture Canada, 662-15374-X.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geosci. Model Dev., 8, 1991-2007, doi:10.5194/gmd-8-1991-2015.

De Bruin, S., Stein, A., 1998. Soil–landscape modeling using fuzzy c-means clustering of attribute data derived from a Digital Elevation Model (DEM). Geoderma 83, 17– 33.

De Gruijter, J.J., Walvoort, D.J.J., van Gaans, P.F.M., 1997. Continuous soil maps—a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma 77, 169– 195.

Dobarco, M. R., Bourennane, H., Arrouays, D., Nicolas P. A. S., Cousin, I., Martin, M. P., 2019. Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma 344, 14–30

Dobos, E., Bialko, T.; Micheli, E.; Kobza, J., 2010. Chapter 25. Legacy soil data harmonization and database development. in J. L. Boettinger et al. (eds), Digital Soil Mapping, Progress in Soil Science, Springer Science and Business Media B. V p. 323-333. DOI 10.1007/978-90-481-8863-5_26

Dunn, O. J., 1964. Multiple comparisons using rank sums. Technometrics 6:241--252

Gallant, J.C., Dowling, T.I. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research, 39(12), 1347.

Gao, Q., Zribi, M., Escorihuela, M. J., Baghdadi, N. 2017. Synergetic use of Sentinel-1 and Sentinel-2 data for doil moisture mapping at 100 m resolution. Sensors 17, 1966; doi:10.3390/s17091966

Geng, X. Y., Dillabaugh, C., Mitchell, S., McNairn, H., Zhu, A. X., Jiao, X. F., Shang, J. L., 2010a. Soil property retrieval using polarimetric RADARSAT-2 imagery. Canada. 4th Global Workshop on Digital Soil Mapping, Rome, Italy.

Geng, X. Y., Fraser, W., VandenBygaart, B., Smith, S., Waddell, A., You, J., and Patterson, G. 2010b. Chapter 26. Toward digital soil mapping in Canada. in J. L. Boettinger et al. (eds), Digital Soil Mapping, Progress in Soil Science 2, Springer Science and Business Media B. V p. 323-333. DOI 10.1007/978-90-481-8863-5_26

Glinka, K. D., 1927. The great soil groups of the world and their development. (Translated from German by C. F. Marbut) Edwards Bros., Ann Arbor, MI.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152:195-207.

Grunwald S J A, Thompson J L, Boettinger J L. 2011. Digital soil mapping and modeling at continental scales: Finding solution for global issues. Soil Science Society of America Journal, 75, 1201–1213.

Hajj, M. E., Baghdadi, N., Zribi, M., Bazzi, H. 2017. Synergic use of Sentinel-1 and Sentinel-2 images for operational soil moisture mapping at high spatial resolution over agricultural areas. Remote Sens. 2017, 9, 1292; doi:10.3390/rs9121292

Harris, J. R., He, J., Rainbird, R., Behnia, P., 2014. A comparison of different remotely sensed data for classification lithology in Canada's arctic: application of the robust classification method and Random Forests. Geoscience Canada 41 (4): 557–584

Hastie, T., Tibshirani, R., Friedman, J., 2009.The Elements of Statistical Learning. Springer Series in Statistics

Hartemink, A.E., McBratney, A.B., 2008. A soil science renaissance. Geoderma 148, 123–129.

He, J., Geng, X., VandenBygaart, B., Martin, T., Hengl, T., MacMillan, R., Shaw, C., 2017. Validation of 250m soil grids in Canada. Pedometrics 2017, Wageningen, Netherland

Heaton, J. 2005. Introduction to neural networks with Java. Heaton Research Inc.

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

Hengl T., Heuvelink G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., et al. 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS ONE 10(6): e0125814. doi:10.1371/journal.pone.0125814

Hengl, T., MacMillian, R. A. 2019. Predictive soil mapping with R. OpenGeoHub foundation. Wageningen, the Netherlands, 370 pages, www.soilmapper.org, ISBN: 978-0-495-30635-0

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B.M., Ruiperez Gonzalez, M., Kilibarda, M., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: global gridded soil information based on Machine Learning. PLOS ONE 12(2): e0169748

Hengl, T., Nussbaum, T., Wright, M. N., Heuvelink, G. B. M. 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables Peer J

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214–215, 141 – 154.

Heung, B., Hodúl, M., Schmidt, M. G., 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma, 290, 51-68

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 62-77

Heuvelink, G.B.M., 2014. Uncertainty Quantification of GlobalSoilMap Products. In: Arrouays, D., NJ, McKenzie, Hempel, J.W., Richer de Forges, A.C., McBratney, A.B. (Eds.), GlobalSoilMap. Basis of the Global Soil Information System. Taylor & Francis, CRC press, Oxon, pp. 335–340.

Heuvelink, G.B.M.; Webster, R. 2001. Modelling soil variation: past, present, and future. Geoderma 100: 26

Hijmans, R. J, 2015. raster: Geographic Data Analysis and Modeling. R package version 2.4-18. http://CRAN.R-project.org/package=raster

Hollander, M., Wolfe, D. A., 1973. Nonparametric Statistical Methods. New York, John Wiley & Sons pp:115–120

Huang, C., Davis, L. S., Townshend, J. R. G., 2002. An assessment of support vector machines for land cover classification. International Journal of Remote Sensing 23 (4): 725–749. doi:10.1080/01431160110040323

Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Sci. Soc. Am. J. 56:836–841.

Iqbal, J., Thomasson, J.A., Jenkins, J.N., Owens, P.R., Whisler, F.D., 2005. Spatial variability analysis of soil physical properties of alluvial soils. Soil Science Society of America Journal, 69 (4), pp. 1338-1350.

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill, New York.

Jenny H., 1994. Factors of soil formation: a system of quantitative pedology. Dover books on Earth sciences. Dover Publications

Jiang, Y., Nishimura, P., van den Heuve, M.R., MacQuarrie, K.T.B., Crane, S., Xing, Z., Raymond, B.G. and Thompson, B. L., 2015. Modeling land-based nitrogen loads from groundwater-dominated agricultural watersheds to estuaries to inform nutrient reduction planning. Journal of Hydrology 529, 213-23

Karthikeyan, L., Pan, M., Wanders, N., Kumar, D. N., Wood, E. F., 2017. Four decades of microwave satellite soil moisture observations: Part 1. A review of retrieval algorithms. Advances in Water Resources 109:106–120

Kavzoglu, T., Colkesen, I., 2009. A kernel functions analysis for support vector machines for land cover classification. International Journal of Applied Earth Observation and Geoinformation 11: 352 – 359. doi:10.1016/j.jag.2009.06.002.

Kempen, B., Brus, D. J., Heuvelink, G. B. M., and Stoorvogel, J. J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinormial logistic regression approach. Geoderma 151:311-326

Kass, G. V., 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29:119-127

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

Kheir, R. B., Greve, M. H., Bocher, P. K., Greve, M. B., Larsen, R., McCloy, K., 2010. Predicative mapping of soil organic carbon in web cultivated lands using classification-tree based models: the case study of Denmark. Journal of Environmental Management 91:1150-1160

Kidd D, Malone B, McBratney A, Minasny B, Webb M. 2015. Operational sampling challenges to digital soil mapping in Tasmania, Australia. Geoderma Reg. 4: 1–10.0.

Knight, J.F., Tolcser, B.P., Corcoran, J.M., Rampi, L.P., 2013. The effects of data selection and thematic detail on the accuracy of high spatial resolution wetland classifications. Photogramm. Eng. Remote Sens. 79, 613–623.

Kroetsch, D.J., Geng, X., Chang, S. X. and Saurette, D. D. 2011. Organic order - Part 1. Wetland Organic soils. Canadian Journal of Soil Science 91:807-822

Kuhn, M., 2019. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R CoreTeam, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan.. caret: Classification and Regression Training. R package version 6.0-64. http://CRAN.R-project.org/package=caret

Kullback, S., & Leibler, R. A., 1951. On information and sufficiency. The Annals of Mathematical Statistics, 22, 79-86.

Lagacherie, P., 2005. An algorithm for fuzzy pattern matching to allocate soil individuals to pre-existing soil classes. Geoderma 128:274–288.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., Saby, N. P. A., 2019. How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma 337:1320-1328

Lagacherie, P., and McBratney, A. B., 2007. Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. p.3–22. In P. Lagacherie et al. (ed.) Digital soil mapping: An introductory perspective. Elsevier, New York.

LeCun, Y., Bengio, Y., Hinton, G. 2015. Deep learning. Nature, 521:436-444, doi:10.1038/nature14539

Li, J. and Chen, W. 2005. A rule-based method for mapping Canada's wetlands using optical, radar and DEM data, International Journal of Remote Sensing, 26:22, 5051-5069, DOI: 10.1080/01431160500166516

Lillesand, T., M. and Kiefer, R. W., 2000. Remote sensing and image interpretation (4th edition). John Wiley and Sons, ISB 0-417-25515-7, pp. 569-570

Lindsay, J. B. 2016. Whitebox GAT: A case study in geomorphometric analysis. Computers & Geosciences, 95: 75-84. DOI: 10.1016/j.cageo.2016.07.003

Liu, F., Geng, X, Y., Zhu, A. X., Fraser, W., Waddell, A., Fitzmaurice J., and Qi, F., 2012. Soil texture mapping over low relief area using land surface feedback dynamic patterns extracted from MODIS. Geoderma, 171:44-52

Liu, F., Geng, X., Zhu, A.,, Walter, F., Song, X., Zhang, G., 2016. Soil polygon disaggregation through similarity-based prediction with legacy pedons. Journal of Arid Land, 8(5): 760–772. doi: 10.1007/s40333-016-0087-7

Liu, J., Pattey, E., Nolin, M. C., Miller, J. R., Ka, O., 2008. Mapping within-field soil drainage using remote sensing, DEM and apparent soil electrical conductivity. Geoderma, 143 (3-4), pp. 261-272.

MacDougall, J.I., C. Veer, and F. Wilson. 1988. Soils of Prince Edward Island: Prince Edward Island soil survey. Agriculture Canada, Supply and Services Canada, Ottawa.

MacMillan, R.A., D.E. Moon, and Coupe, R. A., 2007. Automated predictive ecological mapping in a Forest Region of B.C. Canada: 2001–2005. Geoderma 140:353–373.

MacMillan, R.A., W.W. Pettapiece, S.C. Nolan, and Goddard, T. W., 2000. Ageneric procedure for automatically segmenting landforms into land form elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets Syst. 113:81–109

Malone, B., Minasny, B., Colby, B., 2019. Some methods to improve the utility of conditioned Latin hypercube sampling. PeerJ. 7. e6451. 10.7717/peerj.6451.

Malone, B. P., Minasny, B., Odgers, N. P., McBratney, A. B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232-234, 34-44

Markert B. 2007. Quality assurance of plant sampling and storage. In Quevauviller P (ed.) Qualilty Assurance in Environmental Monitoring: Sampling and Sample Preatreatment. Wiley-VCH Verlag GmbH, Weinheim.

Matheron, G., 1971. The Theory of regionalized variables and its applications. Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu, Paris. Mazaheri, S. A., Koppi, A. J., and McBratney, A. B., 1995. A fuzzy allocation scheme for the Australian Great Soil Groups Classification system. European Journal of Soil Science, December 46:601-612

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

McBratney, A.B., and Odeh, I.O.A., 1997. Application of fuzzy sets in soil science: Fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma, 77:85–113.

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

McKenzie, N. J. and Austin, M. P., 1993. quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57:329-355

McLeod, A. I. and Xu, C., 2018. Best subset GLM and regression utilities. https://cran.r-project.org/web/packages/bestglm/bestglm.pdf

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999

Meinshausen, N., Schiesser, L., 2015. Quantregforest: Quantile Regression Forests. R package. https://cran.r-project.org.

Mendonca-Santos, M.L. and dos Santos, H.G., 2007 The state of the art of Brazilian soil mapping and prospects for digital soil mapping. Pg. 39-45. In: P. Lagacherie, A.B.

McBratney and M. Voltz. Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science – Volume 31 Elsevier, Amsterdam.

Merzouki, A., McNairn, H., Pacheco, A. 2011. Mapping soil moisture using RADAAT-2 data and local autocorrelation stastistics. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 4(1):128:137

Millard, K., Richardson, M., 2013. Wetland mapping with LiDAR derivatives, SAR polarimetric decompositions, and LiDAR–SAR fusion using a random forest classifier. Can. J. Remote. Sens. 39, 290–307

Minasny, B., Berglund, Ö., Connolly, J., Hedley, C., de Vries, F., Gimona, A., Kempen, B., (...), Widyatmanti, W. 2019. Digital mapping of peatlands – A critical review. Earth-Science Reviews, 196, art. no. 102870. Cited 3 times. http://www.sciencedirect.com/science/journal/00128252
doi:10.1016/j.earscirev.2019.05.014

Minasny, B., McBratney, A.B., 2015. Digital soil mapping: a brief history and some lessons. Geoderma. http://dx.doi.org/10.1016/j.geoderma.2015.07.017

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

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388

Mondal, A., Kundu,S., Chandniha, S. K., Shukla, R., Mishra, P. K., 2012. Comparison of support vector machine and maximum likelihood classification technique using satellite imagery. International Journal of Remote Sensing and GIS 1 (2): 116–123

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57, 443-452

Moran, C. J. and Bui, E. N., 2002. Spatial data mining for enhanced soil map modeling. Int. J. Geographical Information Science 6(6):533-549

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review." ISPRS Journal of Photogrammetry and Remote Sensing 66: 247–259

Nauman, T. W. and J. A. Thompson 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma, 213:385 – 399

Niang, M. A., Nolin, M. C., Bernier, M., Perron, I. 2008. Potential of C-band multipolarized and polarimetric SAR for soil drainage classification and mapping. 2008 IEEE International Geoscience and Remote Sensing Symposium - Proceedings; Boston, MA; 6 July 2008 through 11 July 2008; Category number CFP08IGA; Code 76978

Nyiraneza, J., Thompson, B., Geng, X., He, J., Jiang , Y., Fillmore, S., and Stiles, K., 2017. Changes in soil organic matter over 18 years in Prince Edward Island, Canada. Can. J. of Soil Science, 97(4): 745-756, https://doi.org/10.1139/cjss-2017-0033

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L. Schaepman, M. E., and Papritz, A. 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. SOIL, 4, 1–22, 2018 https://doi.org/10.5194/soil-4-1-2018

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

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992a. Fuzzy-c-means and kriging for mapping soil as a continuous system. Soil Sci. Soc. Am. J. 56, 1848–1854

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992b. Soil Pattern Recognition with Fuzzy-c-means: Application to Classification and Soil—Landform Interrelationships. Soil Sci. Soc. Am. J. 56, 505–516

Odgers, N. P., W. Sun, A. B. McBratney, B. Minasny, and D. Clifford 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma, 214-215:91–100

Ontario Geological Survey, 2010. Surficial geology of Southern Ontario; Ontario Geological Survey, Miscellaneous Release--Data 128-REV ISBN 978-1-4435-2483-4 [DVD] ISBN 978-1-4435-2482-7 [zip file]

Pennock, D.J., Anderson, D.W., de Jong, E., 1994. Landscape-scale changes in indicators of soil quality due to cultivation in Saskatchewan, Canada. Geoderma 64 (1-2):1-19

Pal, M., 2005. Random Forest classifier for remote sensing classification. International Journal of Remote Sensing 26 (1): 217–222. doi:10.1080/01431160412331269698

Pebesma, E.J., Bivand, R. S. 2005. Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

Pennock, D.J., Zebarth, B.J., De Jong, E., 1987. Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. Geoderma 40 (3-4):297-315

Poggio, L., Lassauce, A., Gimona, A. 2019. Modelling the extent of northern peat soil and its uncertainty with Sentinel: Scotland as example of highly cloudy region. Geoderma 346:63–74

Qi, F., Zhu, A.X., Harrower, M., Burt, J.E., 2006. Fuzzy soil mapping based on prototype category theory. Geoderma 136, 774–787

Quinlan, J. R., 1993. C4.5: Programs for machine learning. Morgan-Kaufman, San Mateo, CA.

R Core Team, 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rich, E., 1987. Artificial intelligence. In: Shapiro, S. (Ed.), Encyclopedia of Artificial Intelligence, vol. 1. Wiley, New York, pp. 9 – 16

Rossiter, D. G. 2018. Past, present & future of information technology in Pedometrics. Geoderma 324:131–137

Sanchez, P. A., Ahamed,S., Carré,F.,Hartemink, A. E., Hempel,J., Huising, J., Lagacherie, P., McBratney, A. B., McKenzie, N. J., Mendonça-Santos, M. L., Minasny, B., Montanarella, L., Okoth, P., Palm, C. A., Sachs, J. D., Shepherd, K. D., Vågen, T-G, Vanlauwe, B., Walsh, M. G., Winowiecki, and Zhang, G. L., 2009. Digital soil map of the world. Science 325(7):681-680

Schmidt, K., Behrens, T., Scholten, T., 2008. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146:138-146

Schut, P., Smith, S., Fraser, W., Geng, X., Kroetsch, D. 2011. Soil landscape of Canada: building a national framework for environmental information. GEOMATICA 65(3):27-45

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modeling 181:1–15

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

Shangguan, W., Hengl, T., de Jesus, J. M., Yuan, H., and Dai, Y., 2017. Mapping the global depth to bedrock for land surface modeling. Journal of Advances in Modeling Earth Systems, 9(1):65-88

Shaw, C., Hilger, A., Filiatrault, M., Kurz, W., 2018. A Canadian upland forest soil profile and carbon stocks database. Ecology 99(4):989-989, https://doi.org/10.1002/ecy.2159

Shoshany, M.; Svoray, T.; Curran, P. J.; Foody, G. M.; Perevolotsky, A., 2000. The relationship between ERS-2 SAR backscatter and soil moisture: generalization from a humid to semi-arid transect. International Journal of Remote Sensing, 21(11):2337-2343

Shi, X., Long, R., Dekett, R., Philippe, J., 2009. Integrating different types of knowledge for digital soil mapping. Soil Sci. Soc. Am. J. 73:1682-1692

Shi, X., Zhu, A., Burt, J. E., Qi, F., and Simonson, D., 2004. A case-based reasoning approach to fuzzy soil mapping. Soil Sci. Soc. Am. J. 68:885-894

Silva, S., Poggere, G., Duarte de Menezes, M., Carvalho, G., Guilherme, L., Curi, N., 2016. Proximal sensing and digital terrain models applied to digital soil mapping and modeling of Brazilian Latosols (Oxisols). Remote Sensing 8.614. 10.3390/rs8080614.

Small, D., 2011.Flattening Gamma: Radiometric Terrain Correction for SAR Imagery IEEE Transactions on Geoscience and Remote Sensing, 49(8):3081-3093

Smith, S., Bulmer, C., Flager, E., Frank, G. and Filatow, D., 2010. Digital soil mapping at multiple scales in British Columbia. Canada. 4th Global Workshop on Digital Soil Mapping, Rome, Italy.

Soil Classification Working Group, 1998. The Canadian Soil Classification System. Agric. and Agri-Food Can. publ. 1646(revised). 187pp.

Tarnocai, C., Kettles, I., Lacelle, B., 2011. Peatlands of Canada. In: Geological Survey of Canada, Open File 6551. vol. 10 Natural Resources Canada, Ottawa, ON

Thompson J A, Roecker S, Grunwald S, Owens P R. 2012. Digital soil mapping: Interactions with and applications for hydropedology. Hydropedology, 1, 664–709.

Thompson, J. A., T. Prescott, A. C. Moore, J. Bell, D. R. Kautz, J. W. Hempel, S. W. Waltman, and C. Perry. 2010. Regional approach to soil property mapping using legacy data and spatial disaggregation techniques. In 19th World Congress of Soil Science, Brisbane Australia. IUSS.

Tso B. and Mather, P. M., 2009. Classification methods for remotely sensed data. CRC Press, Taylor & Francis Group

Tulyakov, S., Zhang, Z., Govindaraju, V., 2008. Comparison of combination methods utilizing T-normalization and second best score model. in CVPR 2008 Workshop on Biometrics.

van de Poll, H.W. 1989. Lithostratigraphy of the Prince Edward Island redbeds. Atlantic Geol. 25, 23–35.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Vereecken, H., 2017. Pedotransfer functions in Earth system science: Challenges and perspectives. Reviews of Geophysics, 55, 1199–1256. https://doi.org/10.1002/2017RG000581

Vaysse K., Lagacherie P., 2015. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France), Geoderma Regional 4:20-30

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma, 291:55-64

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2018. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma, 311:130-142.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D. et al. 2016. A global spectral library to characterize the world's soil. Earth-Science Reviews, 155:198–230

Walter, C., Lagacherie, P., Follain, S., 2007. Integrating pedological knowledge into digital soil mapping. In: Lagacherie P., McBratney, A. B., Voltz, M. (Eds), Digital Soil Mapping an introductory perspective. page 281-300, Elsevier, Amsterdam

Webster, R. 1994. The development of pedometrics. Geoderma, 62:1-15

Weiss, A. (2001, July). Topographic position and landforms analysis. In Poster presentation, ESRI user conference, San Diego, CA (Vol. 200)

Xu, C., McLeod, A. 2009. Linear Model Selection Using the BICq Criterion.Vignette included in bestglm package, The University of Western Ontario

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

Yang, L., Zhu, A. X., Qi, F., Qin, C. Z., Li, B. L., Pei, T.. 2013. An integrative hierarchical stepwise sampling strategy and its application in digital soil mapping. International Journal of Geographical Information Science, 27, 1–23

Zeraatpisheh, M., Ayoubi, S., Brungard, C. W., Finke, P., 2019. Disaggregating and updating a legacy soil map using DSMART, fuzzy c-means and k-means clustering algorithms in Central Iran. Geoderma. 340:249-258.

Zhang, L., Liu F., Song, X., 2017. Recent progress and future prospect of digital soil mapping: A review. Journal of Integrative Agriculture 16(12):2871–2885

Zhao, Z., Chow, T. L., Rees, H. W., Yang, Q., Xing, Z., and Meng, F. R., 2008a. Predict soil texture distribution using an artificial neural network model. Computer Electronics in Agriculture 65(1):36-48

Zhao, Z., Chow, T. L., Yang, Q., Rees, H. W., Benoy, G., Xing, Z., and Meng, F. R., 2008b. Modeling prediction of soil drainage classes based on digital elevation model parameters and soil attributes from coarse resolution soil maps. Can. J. Soil Sci. 88:787-799

Zhu, A., 1997. A similarity model for representing soil spatial information. Geoderma 77:217-242

Zhu, A., 2000. Mapping soil landscape as spatial continua: the neural network approach. Water Resources Research 36(3):663-677

Zhu A. and Band, L. E., 1994. A knowledge-based approach to data integration for soil mapping. Can J. Remote Sens. 20:408–418.

Zhu, A., Band, L., Dutton, B. And Nimlos, T. J., 1996. Automated soil inference under fuzzy logic. Ecological Modelling 90:123-145

Zhu, A.X., Band, L.E., Vertessy, R., Dutton, B., 1997. Derivation of soil properties using a soil land inference model (SoLIM). Soil Science Society of America Journal 61, 523–533.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K. and Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J. 65:1463–1472

Zhu, A. X, Liu, F., Li, B. L., Pei, T., Qin, C. Z., Liu, G. H., Wang, Y. J., Chen, Y. N., Ma, X. W., Qi, F., Zhou, C. H. 2010a. Differentiation of soil conditions over flat areas using land surface feedback dynamic patterns extracted from MODIS. Soil Science Society of America Journal, 74, 861–869

Zhu, A.X., Yang, L., Li, B., Qin, C., English, E., Burt, J.E., Zhou, C, 2008. Purposive sampling for digital soil mapping for areas with limited data. In: Hartemink, A.E., McBratney, A.B., Mendonca Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. InSpringer-Verlag, New York, pp. 233–245.

Zhu, A., Yang, L., Li, B., Qin, C., Pei, T., and Liu, B., 2010b. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma 155:164-174

Appendices

Appendix A Hands-on examples of inference algorithms

A.1 Information gain and Gene index

Sample data:

Sample No	Slope Gradient	A-Horizon pH	A-Horizon C	Soil Type
1	C	High	Low	A
2	C	Low	High	A
3	B	High	Low	B
4	A	Low	Low	B
5	A	High	Low	B
6	A	Low	High	B
7	B	High	High	B
8	C	Low	Low	B
9	C	High	Low	B
10	A	Low	Low	B
11	C	High	High	B
12	A	Low	High	A
13	B	High	Low	B
14	B	Low	High	A
15	A	Lpw	Low	B

Assume at node t there are 15 samples with attributes Slope Gradient class, soil A-Horizon pH, and A-Horizon C (carbon content) for classifying Soil Type.

With the information gain concept, the information entropy ($I_E(t)$) for node t is calculated as:

$$I_E(t) = -\left(\frac{4}{15}\right)\log_2\left(\frac{4}{15}\right) - \left(\frac{11}{15}\right)\log_2\left(\frac{11}{15}\right) = 0.837$$

For the attribute A-Horizon C, the entropy $I_E(t_{\text{A-Horizon C(A)}})$ and $I_E(t_{\text{A-Horizon C(B)}})$ is calculated as follow according to Equation ():

$$I_E\left(t_{A-HorizonC(A)}\right) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 0.5 + 0.5 = 1$$

$$I_E\left(t_{A-HorizonC(B)}\right) = -\left(\frac{1}{9}\right)\log_2\left(\frac{1}{9}\right) - \left(\frac{8}{9}\right)\log_2\left(\frac{8}{9}\right) = 0.352 + 0.151 = 0.503$$

The information gain for A-Horizon C is then calculated as:

$$Gain(t, A-HorizonC) = I_E(t) - \left(\frac{6}{15}\right)I_E\left(t_{A-HorizonC(High)}\right) - \left(\frac{9}{15}\right)I_E\left(t_{A-HorizonC(Low)}\right)$$
$$= 0.837 - 0.4 - 0.302 = 0.135$$

The gain for the Slope Gradient and A-Horizon pH are calculated in the similar fashion as:

$$Gain(t, SlopeGradient) = I_E(t) - \left(\frac{5}{15}\right)I_E\left(t_{SlopeGradient(C)}\right) - \left(\frac{4}{15}\right)I_E\left(t_{SlopeGradient(B)}\right) - \left(\frac{4}{15}\right)I_E\left(t_{SlopeGradient(A)}\right)$$
$$= 0.837 - 0.324 - 0.216 - 0.26 = 0.037$$

$$Gain(t, A-HorizonpH) = I_E(t) - \left(\frac{7}{15}\right)I_E\left(t_{A-HorizonpH(High)}\right) - \left(\frac{8}{15}\right)I_E\left(t_{A-HorizonpH(Low)}\right)$$
$$= 0.837 - 0.276 - 0.509 = 0.052$$

If we want to split the node t use the information gain measure, the highest information gain among the attributes is A-Horizon C with the value of 0.135 will be used for decision tree node splitting.

If the Gini Impurity Index is used as node splitting criteria, the Gini impurity index of the split at node t for attribute X is then computed as:

$$Gini(t,X)=\left(\frac{n_1}{N_t}\right)I_G\left(t_{X(x1)}\right)+\left(\frac{n_2}{N_t}\right)I_G\left(t_{X(x2)}\right)...+\left(\frac{n_r}{N_t}\right)I_G\left(t_{X(xr)}\right)$$

where

r is number of children at node t

n_i is the number of records at child I

N_t is the total number of samples at node t.

The Gini values for attribute A-Horizon C ($I_G(t_{\text{A-Horizon C(high)}})$ and ($I_G(t_{\text{A-Horizon C(low)}})$ are calculated as:

$$I_G\left(t_{A-HorizonC(high)}\right)=1-\left(\frac{3}{6}\right)^2-\left(\frac{3}{6}\right)^2=1-0.25-0.25=0.5$$

$$I_G\left(t_{A-HorizonC(low)}\right)=1-\left(\frac{1}{9}\right)^2-\left(\frac{8}{9}\right)^2=1-0.013-0.79=0.97$$

$$Gini(t,A-HorizonC)=\left(\frac{6}{15}\right)I_G\left(t_{A-HorizonC(high)}\right)+\left(\frac{9}{15}\right)I_G\left(t_{A-HorizonC(low)}\right)$$

$$=0.2+0.118=0.318$$

Similarly, the Gini impurity index for attribute Slope Gradient and A-Horizon pH are calculated as:

$$Gini(t, SlopeGradient) = \left(\frac{5}{15}\right) I_G\left(t_{SlopeGradient(C)}\right) + \left(\frac{4}{15}\right) I_G\left(t_{SlopeGradient(B)}\right) + \left(\frac{6}{15}\right) I_G\left(t_{SlopeGradient(A)}\right)$$
$$= 0.16 + 0.1 + 0.111 = 0.371$$

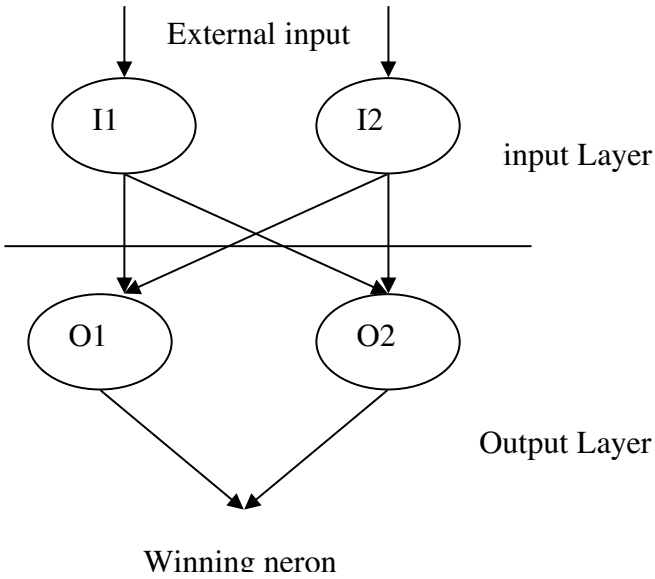
$$Gini(t, A - HorizonpH) = \left(\frac{7}{15}\right) I_G\left(t_{A-HorizonpH(high)}\right) + \left(\frac{8}{15}\right) I_G\left(t_{A-HorizonpH(low)}\right)$$
$$= 0.25 + 0.114 = 0.364$$

Among the three indices, the attribute A-Horizon C (carbon content) has the smallest Gini impurity index and it is the first attribute to be split at the node t. For other decision tree programs, those above examples demonstrated the fundamental concept of decision tree theory although some modifications are conducted surrounding the core concept.

A.2 Example of Kohonen neural network processes

A typical ANN has input, hidden and output layers. However Hope Field neural network only has one layer; Kohonen neural network has two layers of input and output layer. The principles of how a neural network operates are common; all neural networks have neurons and neurons need may be fired or activated when a network is trained and ready to work. Below is just a simple example for a Kohonen neural network to demonstrate how a neural network works (Heaton, 2005).

Assume there is a Kohonen neural network with on two neurons in the input and output layer since we do not expect a hidden layer in the Kohonen neural network structure (Figure 5).



The input neurons are each given floating point numbers that make up for input pattern to the network. A Kohonen neural network requires that these inputs be normalized to the range between -1 and 1. With input pattern to the neural network, we expect that it will react and produce output. Different form feed-forward neural network, Kohonen neural only produce one output values; in another word, when input neurons are presented to a Kohonen neural network, only one single neuron chosen as output neuron. The output from the Kohonen neural network is usually the index of the neuron that fired.

Let's examine how the Kohonen neural network processes information. Assume the input given to the two input neurons is found in Table A2-1

Table A2-1 Example of two neurons

Input Neuron 1(I1)	0.5
Input Neuron 2(I2)	0.75

We also need to know the connection weights between the neurons. These connection weights are given in Table A2-2 and will be discussed later.

Table A2-2 Example of a set of connection weights of the neurons

I1->O1	0.1
I2->O1	0.2
I1->O2	0.3
I2->O2	0.4

As mentioned before, Kohonen neural network requires that its input be normalized. The input to the Kohonen neural network must be between the values -1 and 1. The normalization is often conducted by first calculating the "vector length" of the input data

or vector. This is done by summing the squares of the input vector. For this example, the vector length is calculated as

$$(0.5*0.5) + (0.75*0.75) = 0.8125$$

The normalization factor is the reciprocal of the square root of the vector length.

$$\text{Normalization Factor} = \frac{1}{\sqrt{0.8125}} = 1.1094$$

To calculate the output, the input vector and neuron connection weights must be both considered. The output is the "do product" of the input neurons and their connection weights.

For the first output neuron calculated and normalized as

$$|0.5 \;\; 0.75|.|0.1 \;\; 0.2| = (0.5*0.75) + (0.1*0.2) = 0.395$$

$$0.395*1.1094 = 0.438213$$

For the second output neuron

$$|0.5 \;\; 0.75|.|0.3 \;\; 0.4| = (0.5*0.75) + (0.3*0.4) = 0.495$$

$$0.495*1.1094 = 0.5491531$$

Next step is to map the normalized data into bipolar system. In the bipolar system, the binary zero maps to -1 and binary one remains as 1. Since the input to the neural network normalized to this range, we much also normalize the output of the neurons to this range before we can choose a final or winner output neuron. To conduct the bipolar mapping, we

simply multiple a value by two and subtract one. So the two output values are mapped as below.

0.438213*2 – 1 = -0.123574

0.5491531*2-1 = -0.901694

The first neuron has an output value of -0.123574 and the second one has output value of -0.901694 which beats the neuron two's output of -0.123574.

Now we need to evaluate how the connection weight influences the output. The connection weight can be adjusted to produce more desired output. This modification process of the connection weights is also called training. Just like training for other types of neural networks, the Kohonen neural network is trained by repeating epochs until the calculated error is below an acceptable level. If it is determined that the network training is to be completed, the weights will be re-assigned with random values. During a training cycle, individual epochs are important. Within each epoch, the connection weights are adjusted with specified learning rate and additive or subtractive algorithms.

Learning rate

The learning rate is a constant and need to be specified for the learning algorithms. The learning rate must be a positive number less than 1. In practice, the learning rate is set with a number as 0.4 or 0.5. If the learning rate is higher, the training process goes faster. Too large a learning rate may cause the network to never converge. Learning rate is one of the variables used to adjust connection weight.

Adjusting weights

The entire memory of the Kohonen network is carried by the connection weight between the input and output layer. The weights are adjusted in each epoch. An epoch is identified when training data is presented to the Kohonen network and the connection weights are

adjusted due to the input. The aim of adjusting the connection weight is to produce more favorable outputs even with the same training data. The weight adjustment will be ceased when further weight adjustment doesn't improve the outcomes of the network.

Kohonen proposed an additive method uses following equation for connection weight adjustment.

$$W^{t+1} = \frac{w' + \alpha x}{\| w' + \alpha x \|}$$

Where

x – variable vector that is provided to the neural network

w_t – weight of winning neuron

The double vertical bars represent the vector length used in previous section

W^{t+1} – new connection weight.

The additive weight adjustment works well in most of Kohonen neural network cases. When it doesn't converge, subtractive method as indicated below may be used.

$$e = x - w'$$

$$W^{t+1} = w' + \alpha e$$

A.3 How simple fuzzy set is applied (based on personal communication with Robert MacMillan, 2009)

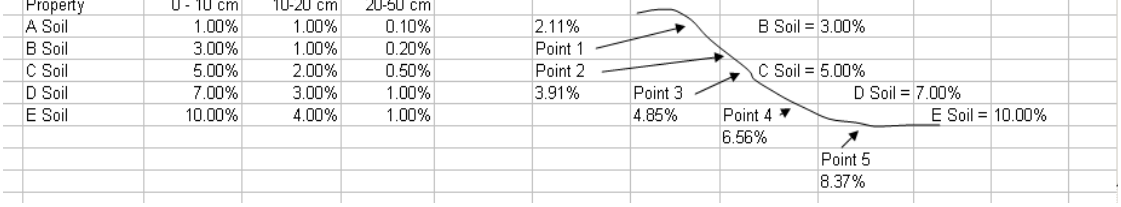
Grid Cell Values	Poiint 1	Point 2	Point 3	Point 4	Point 5
A likelihood	0.600	0.050	0.020	0.010	0.010
A Soil_Value	1.00%	1.00%	1.00%	1.00%	1.00%
A_Soil_Prop	0.60%	0.05%	0.02%	0.01%	0.01%
B likelihood	0.300	0.549	0.200	0.050	0.020
B Soil_Value	3.00%	3.00%	3.00%	3.00%	3.00%
B_Soil_Prop	0.90%	1.65%	0.60%	0.15%	0.06%
C likelihood	0.060	0.300	0.625	0.300	0.070
C Soil_Value	5.00%	5.00%	5.00%	5.00%	5.00%
C_Soil_Prop	0.30%	1.50%	3.13%	1.50%	0.35%
D likelihood	0.030	0.100	0.150	0.500	0.350
D Soil_Value	7.00%	7.00%	7.00%	7.00%	7.00%
D_Soil_Prop	0.21%	0.70%	1.05%	3.50%	2.45%
E likelihood	0.010	0.001	0.005	0.140	0.550
E Soil_Value	10.00%	10.00%	10.00%	10.00%	10.00%
E_Soil_Prop	0.10%	0.01%	0.05%	1.40%	5.50%
Total Likelihood	1.000	1.000	1.000	1.000	1.000
Weighted Mean	2.11%	3.91%	4.85%	6.56%	8.37%

Depth	A Soil	B Soil	C Soil
0-10 cm	1.00%	3.00%	5.00%
10-20 cm	1.00%	1.00%	2.00%
20-50 cm	0.10%	0.20%	0.50%
etc			

Depth Functions

Soil Carbon "Representative Value"

Property	0 - 10 cm	10-20 cm	20-50 cm
A Soil	1.00%	1.00%	0.10%
B Soil	3.00%	1.00%	0.20%
C Soil	5.00%	2.00%	0.50%
D Soil	7.00%	3.00%	1.00%
E Soil	10.00%	4.00%	1.00%



Appendix B

B.1 Example R code of machine learning

```
#' This example shows how to calibrate machine learning models(RF, C5, SVM, and NN)
#' and generate predictions using R packages like caret and raster
#'
#' @param
#' training data: points in ESRI Shapefile
#' covariates: raster data in GTiff format
#' @return
#' prediction of classes: predictions are saved in a GTiff file
#'

library(rgdal)
library(randomForest )
library(raster)
library(caret)
library(plyr)


setwd("D:/PEI_maple_plains_watershed")

# load convariates
cov.names <- list.files("./data", pattern = "tif$", full.names = T)
cov.names
cov <- stack(cov.names)
names(cov)
plot(cov)

# load training data from random sampling
training.pt <-  readOGR(dsn = "./data", layer = "training_sample_random", stringsAsFactors = F)
names(training.pt)
table(training.pt$soiltype)

# check CRS of covariates and training data.
proj4string(cov)
proj4string(training.pt)
# covariate values at training data locations
cov.training.ext <- raster::extract(cov, training.pt)
# add soil type information
cov.training.ext <- cbind.data.frame(training.pt$soiltype, cov.training.ext)
names(cov.training.ext)
names(cov.training.ext)[1] <- "soiltype"
if (!is.factor(cov.training.ext$soiltype)){
  cov.training.ext$soiltype <- as.factor(cov.training.ext$soiltype)
}

# random forest----
rf0 <- train(soiltype~., data = cov.training.ext, method = "rf")
rf0
varImpPlot(rf0$finalModel) # variable importance plot
```

```
print(rf0$finalModel)
predict(cov, rf0, filename = "./result/rf_prediction", format = "GTiff",  progress = "text",  na.rm = T,  type
= "raw",  overwrite = T)

# c5 using caret ------------
str(cov.training.ext)
c50 <-  train(x = cov.training.ext[,-1], y = cov.training.ext[, 1], metric = "Accuracy", method = "C5.0",
verbose = FALSE)
c50
plot(c50)
predict(cov, c50, filename = "./result/c50_prediction", format = "GTiff", progress = "text",na.rm = TRUE,
type = "raw", overwrite = T)

# svm using caret --------------
svmgrid <- expand.grid(sigma = seq(0, 1, 0.1), C = 2^c(0:9))
svm0 <- train(cov.training.ext[,-1], cov.training.ext[, 1], method = "svmRadial", preProc = c("center",
"scale"), metric = "Accuracy", tuneGrid = svmgrid)
svm0
plot(svm0)
varImp(svm0)
svm0$finalModel
predict(cov, svm0, filename = "./result/svm_prediction", format = "GTiff",  progress = "text", na.rm =
TRUE, type = "raw", overwrite = T)

# neural network using caret
nngrid <- expand.grid(.decay = c(0.001, 0.01, 0.1), .size = seq(1, 27, by = 2))
nn0 <- train(cov.training.ext[,-1], cov.training.ext[, 1], method = "nnet", tuneGrid = nngrid, maxit = 1000)
nn0
plot(nn0)
predict(cov, nn0, filename = "./result/nn_prediction", format = "GTiff", progress = "text", na.rm = TRUE,
type = "raw", overwrite = T)
```