QUANTIFYING SPATIAL VARIABILTY OF SOIL TEXTURE IN A GEORGIA PIEDMONT FLOODPLAIN

by

DERRICK PLATERO

(Under the Direction of Matthew Levi and Nandita Gaur)

ABSTRACT

Accurate assessment of spatial variability of soil texture is a significant component of agriculture and environmental modeling. Current soil maps lack detail necessary for intensive management like precision agriculture. Determining optimal sample sizes for creating detailed soil maps is challenging because it is cost and labor prohibitive. In this work, random forest models of soil texture were developed using an 80/20 split for training and testing data, respectively, for 50 iterations of sample sizes between 10-65. Sixty-nine samples were taken from a 40-acre crop field in July 2020 and May 2021 at 0-10, 10-40, 40-70, and 70-100 cm and combined with topographic covariates, electromagnetic conductivity (EM31), and spectral reflectance data as predictors. R^2 and root mean square error (RMSE) varied by soil property and depth. A sample size of 35-45 samples represented the variability of soil texture most depth increments based on the trends in R^2 and RMSE.

INDEX WORDS: Soil, soil texture, random forest, topographic covariates, model, digital soil mapping, precision agriculture

Abbreviations: DEM, Digital Elevation Model; TPI_Landform, Topographic Position Index Landform; PCA, Principal Component Analysis; LSF, Length Slope Factor; SWI, Saga Wetness Index; ECa, Electrical Conductivity; FPL, Flow Path Length; MRRTF, Multiresolution Index of Ridgetop Flatness; MRVBF, Multiresolution Index of Valley Bottom Flatness

QUANTIFYING SPATIAL VARIABILTY OF SOIL TEXTURE IN A GEORGIA PIEDMONT FLOODPLAIN

by

Derrick Platero

BS, New Mexico State University, 2019

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2022

QUANTIFYING SPATIAL VARIABILTY OF SOIL TEXTURE IN A GEORGIA PIEDMONT FLOODPLAIN

by

DERRICK PLATERO

Co-Professors:	Matthew Levi
	Nandita Gaur
Committee:	Daniel Markewitz
	Dorcas Franklin

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2022

DEDICATION

To mom, dad, Trixie (my puppy), Flower (cat) and a past teacher from high school, Ms. Rose, who pushed me to continue my education. Mom and dad for always being supportive and encouraging all the time. Trixie for always being happy to see me and comforting me 24/7. Flower for always giving me the judgmental looks and snuggles. With Ms. Rose, I do remember her giving me advice in high school and what is right and wrong. She was also a mentor.

Winning every race, you just may not; What's important is giving all you got. In running the race, you must never give up; The will to go on, never let anything interrupt.

Millard Lowe

ACKNOWLEDGEMENTS

I would like to thank Dr. Levi and Dr. Gaur for all their help, patience, and guidance throughout my Master's program. Their combined experience and knowledge have helped me with my academic research and writing. I would like to thank Matthew Thibodeaux, Chandler Gruener, Rajneesh Sharma, Therese Thompson, Cortney Stevenson, and Maria Tancredi for their help in the field and laboratory. They have helped me a lot and much appreciated. Finally, I want to thank my family and friends for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Importance of soil texture: A Landscape Scale perspective

Soil serves a major role in facilitating water and nutrient movement in landscapes. Detailed characterizations of soil physical properties are critical inputs for modeling landscape-scale water table fluctuations and surface runoff. Soil properties also determine erodibility potential of the land surface (Wischmeier and Mannering, 1969). Hence, an accurate assessment of spatial variability of soil properties is a significant component of agriculture and environmental modeling (Shit et al., 2016). Presently, application of high spatial resolution sensors and multifaceted modeling can be combined to develop high spatial-resolution soil maps (Robinson et al., 2017). Such maps can improve our ability to manage soils within landscapes.

Soil texture is a physical property of soil representing the relative proportions of particle sizes for a given soil (i.e., sand, silt, and clay). It is one of the most important physical properties of soils and directly affects other critical properties, including susceptibility to erosion, drainage, water-holding capacity, organic matter content, the capacity for leaching nutrients and pollutants and engineering properties (Stevens, 1992; Adugna, 2018). This abiotic factor also influences the distribution of minerals, organic matter retention and microbial biomass (Najmadeen et al., 2010) which explains why it is one of the key components for assessing soil quality and the sustainability of agricultural management practices (Hassink et al., 1993; Villas-Boas et al., 2016). Spatial distributions of soil properties like texture can vary significantly over short distances which can influence

water movement, plant productivity, and erosion at field and landscape scales. Ciampalini et al. (2012) developed a model for analyzing agricultural landscape evolution and suggested that surface soil is currently experiencing rapid evolution due to changes made by climate and humans. Human-imposed changes include those implemented by farmers and policy decision makers across a range of spatial scales (Verburg et al., 2002; Rounsevell et al., 2005; Claessens et al., 2009; Ciampalini et al., 2012) while climate-induced changes relate to deviations in seasonal distribution of climate factors and in the frequency of extreme events predicted by future climate changes (Bernstein et al., 2008; Ciampalini et al., 2012). Therefore, knowledge of spatial variation in soil texture is necessary for sustainable soil management (Pahlavan-Rad and Akbarimoghaddam, 2018) and soil security (McBratney et al., 2014).

Soil texture plays an important role in nutrient availability for crops and other plants. In particular, particle size strongly influences surface area and subsequently cation exchange capacity (CEC) such that clayey soils tend to hold more nutrients than soils with more sand (Seybold et al., 2005). Najmadeen et al. (2010) collected 25 samples of soil at 9 locations in Iraq and analyzed exchangeable cations. Each sample location was random, and collected at a 15 cm depth using a 2.5 cm diameter soil auger (Najmadeen et al., 2010). From the nine collected soils, six different textural soil classes (sandy loam, loamy sand, silty loam, silty clay loam, loam, and clay loam) were detected. The results of soil textures showed no significant effects on concentration of (Mg^{+2} & Na^{+}) ions and available phosphorus. However, high significant differences in Ca^{+2}, HCO_3, Cl^{-}, and $CaCO_3$ among different soil textures were recorded. This may vary world-wide but demonstrates the importance of soil texture for nutrient availability.

Soil texture, in addition to impacting ion retention, also impacts water holding capacity of soils (Olorunfemi et al., 2016) and this information is crucial in assessing water requirements for irrigation schedules and for the prediction of probable crop responses to irrigation (Abu-Hamdeh, 2004). Levi (2017) summarized water holding capacity by soil texture class for 75,736 samples from the National Cooperative Soil Survey (NCSS) pedon database. To achieve this, the Rosetta pedotransfer function (PTF) was used to predict water retention for NCSS samples in the database. Rosetta input includes sand, silt, clay, bulk density, and measured values of water content at field capacity and wilting point (Levi, 2017). The study comprehensively evaluated twelve soil textural classes and showed that the water holding capacity was strongly dependent on texture with silty soils having the highest capacity followed by clays and sands. Consequently, irrigation efficiency is also affected to a large degree by the texture of the surface and upper subsoil layers or horizons. When practicing irrigation obligations, texture is the first piece of soil information needed when making irrigation recommendations (Russell, 1980).

The texture of soil is also important in interpreting data from soil sensors that in the age of precision agriculture, have become crucial for providing insights into agricultural processes governing crop growth, carbon storage, soil water, and nutrient use and can help with timely and informed decisions for management practices (Najmadeen et al., 2010; Singh et al., 2020). The calibration and validation of remotely sensed soil moisture products relies on accurate ground data (Rowlandson et al., 2013). For example, soil texture can affect the spatial variability of soil moisture dynamics which needs to be quantified using in-situ or field sensors for several applications. For instance, the relationship between the signal of

electrical conductivity of fields measured using the Electromagnetic conductivity meter (EM) and soil moisture varies with soil texture (Kargas et al., 2013; Singh et al., 2020). Kargas et al. (2013) used disturbed and undisturbed soil cores to examine how soil texture affects EM readings. In the same study, they also evaluated the impact of soil texture (varying from sands to clay) on the calibration of a water content reflectometer sensor, TDR300 (Field Scout TDR300, Product manual., 2022). The influence of soil texture led to variability in the calibration results for water content where sand had the best results while clay was less accurate. A major finding of this study was that the TDR300 had difficultly calibrating, and soil EC affected the permittivity values with levels as low as EC <2 dS/m. The relationship remained linear up to EC 2dS/m that corresponded to bulk soil EC value of 0.6dS/m. EC values >2dS/m was not linear which made the TDR300 calibration problematic. They found higher EC values in clay and it was problematic when there was a change in soil temperature and leaching after irrigation. Similar to Kargas et al. (2013), Ponizovsky et al. (1999) conducted research on TDR and found an increase in variability in clay content resulted in varied moisture content. Their study focused on the influence of quartz sand (particles ranging from 0.05 – 2.00 mm) and sand-kaolin mixtures in Ukraine and forest soil samples from Moscow. The soils were classified as Eutric Podzoluvisols, Orthic Greyzems, and Luvic Chernozems (Unesco, 1987; Ponizovsky et al., 1999). All the samples were taken from upper 20 cm of the A horizon. TDR measurements changed in curves for fine-texture samples in volumetric content of 0.13 to 0.27. The uncertainty in the measured volumetric water content (θv) was higher using the factory calibration root mean square difference compared to the laboratory calibration for the different soil structures and texture classes (Ponizovsky et al., 1999; Kargas et al., 2013; Rowlandson et al., 2013; Singh et al., 2020).

The performance of the TDR was affected by soil texture. Although, the Topp and De Loor equations did not have fitting parameters and are satisfactory models for coarse-textured soils and sand-clay mixtures, in general, the uncertainty in estimation of soil water depth was greater than the uncertainty in estimation of soil water depletion by the sensors installed in the field, and the uncertainties in estimation of depth and depletion were lower using the calibration developed from the undisturbed soil samples (Singh et al., 2020).

1.2 Spatial variability in soil texture

Spatial variability in soil texture differs depending on the pedologic and geologic forming factors (Bockheim et al., 2014) as well as tillage and other management practices (Saleh, 2018). Soil spatial variability refers to soil properties measured at various locations that display different values (Mulla and McBratney, 2001; Wendroth et al., 2011). Like other soil physical properties, texture is highly variable and exhibits scale dependent spatial variability (Wendroth et al., 2011).

For agricultural purposes, some pertinent questions about soil physical properties are: What is the most representative scale of variability and at which scale should we measure? Will this vary by major land resource area? How does variability change with scales? Traditional soil survey relies on the tacit knowledge of a soil scientist to represent the soil-landscape relationship with limited information in a timely manner (Bui, 2004). While these methods provide a great deal of information for a variety of uses, it is often the case that site-specific uses such as precision agriculture require more detailed information (Söderström et al., 2016). For example, Söderström et al. (2016) mention that no detailed general maps for farm level use on topsoil texture are available in Sweden. Each farmer must perform soil sampling and pay for the test analysis. This can be expensive and time consuming (Mallarino

and Wittry, 2000). Traditional methods of soil mapping only cover one third of the land mass on earth and this is for scales finer than 1:1000000 (Hartemink et al., 2013). The traditional maps are based on geomorphic rules of spatial arrangement of soils and at each scale show soil distribution patterns (Gessler et al., 1995; Hartemink et al., 2013). Traditional soil maps are often static and are based on obsolete data (Omuto et al., 2012). Furthermore, their reproducibility is sometimes challenging, as the metadata are may be unfinished, and methods for drawing the maps are often not well documented, as they mainly come from the mental model of the soil surveyor who drew them (Arrouays et al., 2020a). In general, the soil surveyors restrict polygons that are considered as relatively homogeneous while they are based on high categorical levels of various soil classifications and provide limited information about the uncertainty of soil attributes (Terribile et al., 2011; Omuto et al., 2012). Another limitation of traditional soil mapping is that they seldom contain detailed, site-specific information. They are not envisioned for use as primary regulatory tools in site-specific permitting decisions. They are useful for broad regulatory planning and application for multiple uses (Hempel et al., 2008).

The small scale at which soil physical property data is needed for agricultural production and the sustainable use of soil makes it logistically difficult to sample soil properties at the required spatial resolution. It takes time and money to sample soil to produce a good soil map. A useful solution to this problem is to identify easily measurable covariates that can be used to map the spatial variation in soil physical properties. Spatial predictions of physical and hydrological properties with depth at the field-scale are often related to microtopography, which can be represented with detailed topographic indices (Mulla and McBratney, 2001). The subsurface patterns, however, can be quite challenging to identify

with surface reflectance and topographic variables alone. Proximal sensing techniques like electromagnetic induction (EM) are useful for identifying subsurface features associated with changes in ground conductivity (Daily et al., 2004). Digital soil mapping (DSM) has emerged as a powerful tool for such detailed soil mapping. It utilizes relationships between different environmental covariates (e.g., digital elevation models, aspect, ECa, spectral imagery, saga wetness index) and soil properties (e.g., sand, silt, clay, carbon, nutrients) to provide estimates of their spatial distribution.

1.3 Importance of Digital Soil Mapping

High spatial resolution imagery along with DSM has helped bridge the classic theories of soil science with a state-of-the-art computing age to produce high resolution predictive soil maps (Boettinger et al., 2010a). DSM is the generation of geographically referenced soil databases based on quantitative relationships between spatially explicit environmental data and measurements made in the field and laboratory with the intention of predicting soil classes or properties from point data using a statistical algorithm (McBratney et al., 2003a; Scull et al., 2003a). It is a useful approach to predict the spatial variability of soil properties and reduce the need to aggregate soil information based on a set mapping scale (McBratney et al., 2003). This body of work commonly uses Hans Jenny's state factor equation:

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

where cl, o, r, p, and t represent climate, organisms, relief, parent material, and time, respectively (Jenny, 1941). The clorpt framework has been approximated using various environmental spatial data layers to predict soil types and properties based on field

observations, which can be expensive and time consuming to obtain (Amundson and Jenny, 1991). While Jenny's cloprt model is largely conceptual, it is often used quantitatively. This model has been important mainly because it changed the way soils have been studied and leading to the development of more empirical models to describe pedogenesis using a mathematical approach (Ma et al., 2019). Not surprisingly, DSM requires a substantial set of environmental mapping layers to predict soil characteristics (Sanchez et al., 2009a). The approach of DSM has evolved to include the spatial proximity to neighboring soils with the introduction of the scorpan spatial prediction function:

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

where S_c is soil classes and S_a is soil attributes at spatial position x is a function of soil factors (s), climate (c), organisms, which include land use, human effects, and management (o), relief (r), parent materials (p), age or time (a), spatial position (n), and e is the spatially correlated errors (McBratney et al., 2003). The form of f can be a simple linear model to more complicated data-mining tools such as regression trees and random forests (Minasny, 2013). DSM is now recognized as a distinct sub-discipline of soil science (Minasny and McBratney 2015). Soil maps are an effective tool for transmitting information about the spatial distribution of soil attributes (McBratney, 2015) and most work in DSM is based on building numerical models relating field soil observations and some combination of clorpt or scorpan factors (McBratney, 2015).

1.4 Covariate selection for Digital Soil Mapping

Choosing the right set of covariates is one of the most important factors that affects the accuracy of DSM (Liang et al., 2020). According to Liang et al. (2020), the covariate selection knowledge contained in the DSM application is tacit and non-systematic, which means that such knowledge is difficult to formalize in clear rules or mathematical equations. It can be a challenge for users to know the potential for covariates and their practicability for a specific DSM application at hand because the spatial patterns of the same covariate may be different for different locations. This can be an issue with users of DSM who may not have enough knowledge about the details of soil-landscape relationships, such as hydrologists, ecologists, and natural resource managers (Rossiter et al., 2015; Jiang et al., 2016; Dharumarajan et al., 2019; Liang et al., 2020). At present, statistical or machine learning methods are the main approaches applied to assist users in the selection of covariates for DSM (Liang et al., 2020). Correlation plots are good visual aids when selecting covariates because they allow for simple interpretations of how specific soil properties are related to covariates.

Knowledge of general soil forming processes for a given area is very beneficial when deciding which covariates to use, but it is also important to recognize that the availability of covariates is not the same for all areas. Selecting random covariates may result in a modest model performance. Depending on the area and size of the study region, some users apply geological maps, parent material maps, satellite bands, and land use maps for larger areas. These usually have a higher pixel size ranging from 5 m to 300 m. Samuel-Rosa et al., (2015) completed a study in Brazil for a ~ 2000 ha area on the southern edge of the plateau of the Parana Sedimentary Basin, Rio Grande do Sul, Brazil to determine if more detailed environmental covariates deliver more accurate soil maps. Eight continuous predictor variables

representing topography were derived from Light Detection and Ranging (LiDAR), namely slope, aspect, flow accumulation, topographic wetness index, stream power index, topographic position index, and northernness. For categorical predictor variables, land use cover, soil maps, and geologic maps were applied. Soil organic carbon, clay, and effective cation exchange capacity were predicted with R^2 values for clay ranging from 0.460 to 0.485 using the categorical predictor variables. Clay was moderately well predicted using less detailed environmental covariates, with small improvement when using the more detailed covariates. Clay was expected to have a strong correlation with topography and parent material. A notable observation that resulted from their study was that if a low-resolution covariate yields poor predictions, the more detailed version has the possibility to make an improvement in the model performance. But Eldeiry and Garcia, (2008) found that their low R^2 values for soil salinity was due to some locations with high soil salinity. In some areas expected to have little biomass, there were weeds growing that produced a false indication of high biomass when the image is processed. They used Landsat images to help predict their models. Thompson et al., (2001) suggested that model performance will most likely show less improvement using more accurate and detailed covariates if their less detailed version has already made accurate predictions. Samuel-Rosa et al. (2015) concluded that using more detailed covariates resulted in a slight increase in the prediction accuracy of models and choosing whether to use more detailed covariates depends on the strength of the relationship between the covariates and the soil property being modeled. Also, a more precise covariate has a higher potential to improve model prediction when the soil property is poorly predicted by its less detailed version. For example, if topographic position index is used as a variable, then also using watershed metrics, landforms, and slope position may bring the model performance down due to collinearity of the covariates. Collectively these studies support

the idea that covariate resolution is scale dependent and cannot always be transferred across different biomes.

Even though the DSM community is increasingly relying on machine learning to handle the multitude of covariates, the knowledge aspect of traditional soil mapping is still essential. Both approaches capture the elements of the cloprt model (Jenny, 1941) albeit in different ways. A wide variety of covariates are used in soil prediction models to represent terrain, surface reflectance, and climate with a significant reliance on remotely sensed imagery. One of the easiest derived covariates comes from topographic derivatives that are correlated state factors and used as main predictor variables. Ma et al. (2019) have indicated that not all soil state factors have representative covariates that are exactly related to a particular factor and some covariates have implicit or multiple-factor relations. For example, direct estimates of time can be difficult to represent in predictive models unless incorporated manually, however, indirect estimates of time can be inferred from relative landscape position, surface reflectance, weathering indices based on gamma radiometry, or parent material maps (Ma et al., 2019). Even though a researcher may be an expert in the landscape of the study site, the selection of covariates can be biased. To test this hypothesis, Brungard et al. (2015) used three different methods of selecting covariates in their study: 1) expert soil scientist knowledge that used the covariates in the conditioned Latin Hypercube design (cLHS), 2) 113 covariates derived from the digital elevation model (DEM), and 3) those covariates that were selected by recursive feature elimination from all available covariates. They found that models using all available covariates were as accurate, or slightly more accurate (higher κ, lower Brier scores), than models using the covariates selected by soil scientists for each area and for soil class prediction, the covariates selected by recursive elimination gave accurate model results. They also found that the models the soil scientist

selected as covariates had the worst model performance despite being very familiar with the soil-landscape relationships in the area. When selecting potential covariates for soil property prediction, it is best to employ mathematical, statistical, and numerical models to analyze the direct or indirect relationship of soil classes and the environment (Ma et al., 2019). Hengl et al. (2007b) suggested four types of models for soil class mapping.

1. Use of pure classification techniques: These are not interpolators and are classified as images and remote sensing bands.
2. Use of multinomial logistic regression: This is a conventional statistical technique and is recommended to be used where there are more than two classes.
3. Interpolation: Point observations to interpolate soil categorical variables are a useful technique
4. Expert knowledge: Use of expert knowledge to work on data preparation of the soil-landscape combinations.

1.5 Selection of topographic covariates

Topography is one of the soil-forming factors as seen in the Hans Jenny clorpt equation (Jenny, 1941) and directly or indirectly controls the spatial distribution of physical, chemical, and biological soil properties (Florinsky, 2012). It influences soil properties through two main physical processes: the gravity-driven lateral migration and accumulation of water and spatial differentiation of the temperature regime of slopes (Florinsky, 2016). The influence of topography on soil properties can also be influenced by the management or tillage practice in an agricultural setting.

Li et al. (2020) assessed the predictive power of topographic covariates to estimate soil properties and processes at Walnut Creek Watershed, Iowa. They examined two types of

topographic covariates used in soil property modeling. i.e., primary and secondary according to calculation methods. Primary covariates are calculated from elevation and slope, aspect, and curvatures and can be split into two categories, local and nonlocal. Local covariates describe surface geometry at a given point and nonlocal can be computed with a second-order finite difference scheme as relative positions at a selected location (Wilson and Gallant, 2000). Secondary covariates describe the spatial variability in processes such as water content distribution and soil erosion. (Li et al., 2020). Zhu et al. (2010) investigated the use of repeated EMI surveys, in combination with depth to bedrock and terrain attributes, to improve soil mapping in a 19.5-ha agricultural landscape. Results showed that the optimal use of EMI depends on the targeted soil properties, landscape characteristics, specific EMI meter and its setting, and the timing of the survey. A combination of repeated EMI surveys, depth to bedrock, and terrain attributes provided the best mapping of soils in this agricultural landscape and doubled the accuracy of map unit purity compared with the existing second-order soil map. Recent investigations with ground conductivity meters have shown that electrical conductivity measurements using electromagnetic induction have the potential for quick non-invasive soil water content measurement (Sheets and Hendrickx, 1995). Mapping spatial distribution of average soil property using geophysical instruments such as the EM31 for this project will measure bulk soil electrical conductivity and groundwater.

1.6 Scope of Pedometrics and Machine Learning in DSM

Applications of machine learning in DSM have increased rapidly in the last 10 years (Padarian et al., 2020) and soil science utilizes pedometrics to learn and understand how soil is distributed through time and space with data. Pedometrics is a new term coined by A.B. McBratney which stems from two Greek words pedos (soil) and metron (measurement) and is

defined as “the application of mathematical and statistical methods for the study of the distribution and genesis of soils” (McBratney et al., 2018). Machine learning offers many advantages for soil prediction, but selecting an appropriate machine learning method for DSM can be challenging due to the large number of approaches. Khaledian and Miller (2020) discuss machine learning algorithms to identify relationships between soil properties and various covariates across landscapes. They reviewed the number of research papers and books on DSM using machine learning algorithms (Multiple linear regression (MLR), Cubist, Random Forest (RF), k-nearest neighbors (KNN), artificial neural networks (ANN), and support vector regression (SVR) and found the number of citations went from 100 to 2100 involving DSM from 2005 to 2018. All the data was extracted from a keyword search of “digital soil mapping”. DSM has seen a rapid increase in machine learning methods with RF being the most common followed by ANN and MLR (Were et al., 2015; Khaledian and Miller, 2020). The application of RF may be because it uses “bootstrapping” which decreases the variance and improves the stability of the results.

In response to the increasing demand for information on soil properties for environmental modeling, more studies have been shown over the past decade to measure the spatial variability of soil properties on a regional to global scale (Grimm et al., 2008; Hengl et al., 2017; Duchesne and Ouimet, 2021). These investigations rely on the collection of numerous soil field records while developing statistical methods that allow users to calculate consistent and dependable spatial predictions of soil properties at spatial scales (McBratney et al., 2003a; Sanchez et al., 2009b; Duchesne and Ouimet, 2021). The most cutting-edge soil mapping methods involve producing predictions using optimal statistical models that define statistical relationships between observed soil properties and a set of rasterized environmental covariates that are

relevant to explain the distribution of soil properties in the entire area to be mapped (Malone et al., 2017; Hengl and MacMillan, 2019; Duchesne and Ouimet, 2021). Many studies have predicted soil texture for large areas, i.e., 500,000 km^2 or more. Duchesne and Ouimet (2021) used random forest in their study for spatial variability of soil properties. Their study area was 583,000 km^2 in Canada and included 29,570 soil samples. In their statistical modeling, computation of isometric log ratio was done to first transform soil texture fractions and used two values for subsequent statistical modeling. Tree based random forest machine learning algorithms were used to predict and to fine-tune the model and cross validation was used (Duchesne and Ouimet, 2021). The model performance had R^2 values of 0.46 and 0.57 and mean absolute errors of 0.39 and 0.41.

The sample size for DSM is another crucial factor that can also have a strong impact on DSM accuracy. For example, Somarathna et al. (2017) and Long et al. (2018) showed that prediction accuracy of a modelling method is sensitive to sample sizes. It is because sample sizes affect both fitting and prediction of a model (Khaledian and Miller, 2020; Sun et al., 2019). Generally, a model based on more samples is more reliable than one based on fewer samples (Kuang and Mouazen, 2012). However, sample sizes for DSM cannot increase infinitely, due to a limited budget for a soil survey. One example, Lai et al. (2021) had 1861 sample points with a grid of 16 km x 16 km in hilly and mountainous areas. Here they were predicting soil organic carbon in Guangdong, China with 179,700 km^2 as their coverage area. Their bias values reached a threshold of 0.07 at a sample size of 800 followed by a trend of lower R^2 for larger sample sizes. For sample sizes larger than 300, RMSE of each modelling method does not change. Modeling selection is more important when sample size is small, whereas larger sample sizes tend to result in similar performance regardless of the model. However, when sample sizes are

large, e.g., more than 1000, sample sizes have a much greater impact than modelling methods. For coverage probability, the modelling methods have a greater impact than the sample sizes. Coverage probability refers to the probability that an estimated prediction interval covers the corresponding measured value. It measures the accuracy of uncertainty in prediction in terms of variance in prediction. Loiseau et al. (2021) evaluated 8100 points for a total area of 5208 km^2, corresponding to a density of one profile per 0.64 km^2 in the Mayenne region, France. Results for sand showed R^2 values of 0.37 for ordinary kriging and 0.33 for quantile random forest (QFR). R^2 for clay percentages for OK random was 0.27 and QFR was 0.26. As for silt R^2 , OK random was 0.30 and QRF 0.33. The RMSE value for clay was 57% for OK random and QRF. Silt RMSE was 83% for QFR and 80% for OK random. Sand RMSE for OK random was 83% and 89% for QFR. The results showed that, with increasing density of observations, OK performed as well or even better than QRF, depending on the particle-sized fraction. For silt prediction, OK was systematically better than QRF. However, the forecast intervals were much larger for OK than for QRF, and OK did not seem to estimate uncertainty correctly. Overall, the performance indicators increased with the density of observations with a threshold at about 1 profile per 2 km^2 which suggests that the main limitation of DSM prediction accuracy using QRF is the amount of data collected in the field, not the type of calibration sampling strategy. (Loiseau et al., 2021).

1.7 Research Objectives

The objectives for this research are to 1) estimate the optimal sample size for determining spatially distributed soil texture for a field in the floodplain of the Georgia Piedmont and 2) determine the relative importance of different covariates used in the estimation. The optimal sample size was determined by the R^2 and RMSE for sand, silt, and clay. Machine learning techniques specifically, Random Forests (RF) were used to create prediction maps of soil texture based on several environmental covariates.

1.8 References

Abu-Hamdeh, N.H. 2004. The Effect of tillage treatments on soil water holding capacity and on soil physical properties. (669): 6.

Adugna, G. 2018. A review on impact of compost on soil properties, water use and crop productivity. Agric. Sci. Res. J. Vol. 4(3): 93–104. doi: 10.14662/ARJASR2016.010.

Amundson, R., and H. Jenny. 1991. The place of humans in the state factor theory of ecosystems and their soils. Soil Sci. doi: 10.1097/00010694-199101000-00012.

Anderson-Sprecher, R. 1994. Model Comparisons and R 2. Am. Stat. 48(2): 113–117. doi: 10.1080/00031305.1994.10476036.

Arrouays, D., A. McBratney, J. Bouma, Z. Libohova, A.C. Richer-de-Forges, et al. 2020a. Impressions of digital soil maps: The good, the not so good, and making them ever better. Geoderma Reg. 20: e00255. doi: 10.1016/j.geodrs.2020.e00255.

Arrouays, D., L. Poggio, O.A.S. Guerrero, and V.L. Mulder. 2020b. Digital soil mapping and GlobalSoilMap. Main advances and ways forward. Geoderma Reg. 21: e00265. doi: https://doi.org/10.1016/j.geodrs.2020.e00265.

Bernstein, L., P. Bosch, O. Canziani, Z. Chen, R. Christ, et al. 2008. Climate Change 2007 Synthesis report. Intergovernmental Panel on Climate Change.

Biswas, A., and Y. Zhang. 2018. Sampling Designs for Validating Digital Soil Maps: A Review. PEDOSPHERE 28(1): 1–15. doi: 10.1016/S1002-0160(18)60001-3.

Bockheim, J.G., A.N. Gennadiyev, A.E. Hartemink, and E.C. Brevik. 2014. Soil-forming factors and Soil Taxonomy. Geoderma 226–227: 231–237. doi: https://doi.org/10.1016/j.geoderma.2014.02.016.

Boettinger, J.L., D.W. Howell, A.C. Moore, A.S. Hartemink, and S. Kienast-Brown. 2010a. Digital soil mapping; bridging research, environmental application and operation.

Boettinger, J.L., D.W. Howell, A.C. Moore, A.S. Hartemink, and S. Kienast-Brown. 2010b. Digital soil mapping; bridging research, environmental application and operation.

Böhner, J., R. Köthe, O. Conrad, J. Gross, A. Ringeler, et al. 2001. Soil regionalisation by means of terrain analysis and process parameterisation. Eur. Soil Bur.

Böhner, J., and T. Selige. 2006. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. SAGA - Anal. Model. Appl. doi: 10.1186/1471-2288-4-5.

Bruce Robert Russell, 1926-. 1980. Irrigation of crops in the southeastern United States : principles and practice /. Agricultural Research, Southern Region, Science and Education Administration, US Dept of Agriculture, New Orleans, La. (P.O. Box 53326, New Orleans, La., 70153) : Watkinsville, Ga. (P.O. Box 555, Watkinsville, Ga., 30677) :

Brungard, C.W., J.L. Boettinger, M.C. Duniway, S.A. Wills, and T.C. Edwards. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240: 68–83. doi: 10.1016/j.geoderma.2014.09.019.

Bui, E.N. 2004. Soil survey as a knowledge system. Geoderma 120(1): 17–26. doi: https://doi.org/10.1016/j.geoderma.2003.07.006.

Burwell, R.E., D.R. Timmons, and R.F. Holt. 1975. NUTRIENT TRANSPORT IN SURFACE RUNOFF AS INFLUENCED BY SOIL COVER AND SEASONAL PERIODS. Proc Soil Sci Soc Am. doi: 10.2136/sssaj1975.03615995003900030040x.

Chang, D.-H., R. Kothari, and S. Islam. 2003. Classification of soil texture using remotely sensed brightness temperature over the Southern Great Plains. IEEE Trans. Geosci. Remote Sens. 41(3): 664–674.

Ciampalini, R., S. Follain, and Y.L. Bissonnais. 2012. LandSoil: A model for analysing the impact of erosion on agricultural landscape evolution. Geomorphology 175–176: 25–37. doi: https://doi.org/10.1016/j.geomorph.2012.06.014.

Claessens, L., J.M. Schoorl, P.H. Verburg, L. Geraedts, and A. Veldkamp. 2009. Modelling interactions and feedback mechanisms between land use change and landscape processes. Agric. Ecosyst. Environ. 129(1): 157–170. doi: https://doi.org/10.1016/j.agee.2008.08.008.

Conrad, O., B. Bechtel, M. Bock, H. Dietrich, E. Fischer, et al. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. doi: 10.5194/gmd-8-1991-2015.

Dash, P.K., N. Panigrahi, and A. Mishra. 2022. Identifying opportunities to improve digital soil mapping in India: A systematic review. Geoderma Reg. 28: e00478. doi: 10.1016/j.geodrs.2021.e00478.

Dharumarajan, S., R. Hegde, N. Janani, and S.K. Singh. 2019. The need for digital soil mapping in India. Geoderma Reg. 16: e00204. doi: 10.1016/j.geodrs.2019.e00204.

Donovan, M., A. Miller, M. Baker, and A. Gellis. 2015. Sediment contributions from floodplains and legacy sediments to Piedmont streams of Baltimore County, Maryland. Geomorphology 235: 88–105. doi: https://doi.org/10.1016/j.geomorph.2015.01.025.

Duchesne, L., and R. Ouimet. 2021. Digital mapping of soil texture in ecoforest polygons in Quebec, Canada. PeerJ 9: e11685. doi: 10.7717/peerj.11685.

Duffera, M., J.G. White, and R. Weisz. 2007. Spatial variability of Southeastern U.S. Coastal Plain soil physical properties: Implications for site-specific management. Geoderma 137(3): 327–339. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edselp&AN=S0016706106002618&site=eds-live&custid=uga1.

Ferguson, R.B., and G.W. Hergert. 2009. Soil sampling for Precision Agriculture. Soil Sampl. Precis. Agric. https://extensionpublications.unl.edu/assets/pdf/ec154.pdf.

Field Scout TDR300, Product manual. 2022. Spectrum Technologies Inc.: Plainfield, IL.

Florinsky, I.V. 2016. Digital terrain analysis in soil science and geology. Second edition. Academic Press is an imprint of Elsevier.

Franzen, D.W. 2018. Soil Variability and Fertility Management. Precision Agriculture Basics. John Wiley & Sons, Ltd. p. 79–92

Frazier, W. 2006. Geologic regions of Georgia - new Georgia encyclopedia. Geol. Reg. Ga. https://www.georgiaencyclopedia.org/articles/science-medicine/geologic-regions-of-georgia-overview/.

Georgia Soil Survey 136 - Southern Piedmont | NRCS Georgia. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/ga/soils/surveys/?cid=nrcs144p2_021883 (accessed 14 March 2022).

Georgia Weather - Automated Environmental Monitoring Network Page. http://weather.uga.edu/index.php?content=tp&variable=TC (accessed 13 March 2022).

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. 1995. Soil-landscape modelling and spatial prediction of soil attributes. Int. J. Geogr. Inf. Syst. doi: 10.1080/02693799508902047.

Godinho Silva, S.H., P.R. Owens, B.M. Silva, G. César de Oliveira, M. Duarte de Menezes, et al. 2015. Evaluation of Conditioned Latin Hypercube Sampling as a Support for Soil Mapping and Spatial Variability of Soil Properties. Soil Sci. Soc. Am. J. doi: 10.2136/sssaj2014.07.0299.

Grimm, R., T. Behrens, M. Märker, and H. Elsenbeer. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146(1–2): 102–113. doi: 10.1016/j.geoderma.2008.05.008.

Harper, W.V. 2006. Visualization tools to aid in the understanding of geostatistics.

Hartemink, A.E., P. Krasilnikov, and J.G. Bockheim. 2013. Soil maps of the world. Geoderma 207–208: 256–267. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edselp&AN=S0016706113001572&site=eds-live&custid=uga1.

Hartemink, A., and B. Minasny. 2016. Developments in Digital Soil Morphometrics. Digital soil morphometrics /. Springer, Switzerland : p. 425–433

Hartemink, A.E., Y. Zhang, J.G. Bockheim, N. Curi, S.H.G. Silva, et al. 2020. Chapter Three - Soil horizon variation: A review. In: Sparks, D.L., editor, Advances in Agronomy. Academic Press. p. 125–185

Hassink, J., L.A. Bouwman, K.B. Zwart, J. Bloem, and L. Brussaard. 1993. Relationships between soil texture, physical protection of organic matter, soil biota, and c and n mineralization in grassland soils. Geoderma 57(1–2): 105–128. doi: 10.1016/0016-7061(93)90150-J.

Hempel, J.W., R.D. Hammer, A.C. Moore, J.C. Bell, J.A. Thompson, et al. 2008. Challenges to Digital Soil Mapping. In: Hartemink, A.E., McBratney, A., and Mendonça-Santos, M. de L., editors, Digital Soil Mapping with Limited Data. Springer Netherlands, Dordrecht. p. 81–90

Hengl, T., G.B.M. Heuvelink, and D.G. Rossiter. 2007a. About regression-kriging: From equations to case studies. Spat. Anal. 33(10): 1301–1315. doi: 10.1016/j.cageo.2007.05.001.

Hengl, T., and R.A. MacMillan. 2019. Predictive Soil Mapping with R. Lulu.com.

Hengl, T., J. Mendes de Jesus, G.B.M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, et al. 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12(2): 1–

40. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=fsr&AN=121343131&site=eds-live&custid=uga1.

Hengl, T., N. Toomanian, H.I. Reuter, and M.J. Malakouti. 2007b. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. Pedometrics 2005 140(4): 417–427. doi: 10.1016/j.geoderma.2007.04.022.

Heung, B., H.C. Ho, J. Zhang, A. Knudby, C.E. Bulmer, et al. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265: 62–77. doi: 10.1016/j.geoderma.2015.11.014.

Hummel, J.W., L.D. Gaultney, and K.A. Sudduth. 1996. Soil property sensing for site-specific crop management. Comput. Electron. Agric. 14(2–3): 121–136. doi: 10.1016/0168-1699(95)00043-7.

Ike, A.F., and J.L. Clutter. 1968. The Variability of Forest Soils of the Georgia Blue Ridge Mountains. Soil Sci. Soc. Am. J. 32(2): 284–288. doi: 10.2136/sssaj1968.03615995003200020034x.

Jenny, H. 1941. Factors of Soil Formation. Soil Sci. doi: 10.1097/00010694-194111000-00009.

Jiang, J., A.-X. Zhu, C.-Z. Qin, T. Zhu, J. Liu, et al. 2016. CyberSoLIM: A cyber platform for digital soil mapping. Geoderma 263: 234–243. doi: 10.1016/j.geoderma.2015.04.018.

JMP®, J.V. 16. 2021. JMP student version.

Kargas, G., N. Ntoulas, and P.A. Nektarios. 2013. Soil texture and salinity effects on calibration of TDR300 dielectric moisture sensor. Soil Res. 51(4): 330. doi: 10.1071/SR13009.

Kerry, R. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. Geoderma. 140(4): 383–396.

Khaledian, Y., and B.A. Miller. 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. 81: 401–418. doi: https://doi.org/10.1016/j.apm.2019.12.016.

Kidd, D., R. Searle, M. Grundy, A. McBratney, N. Robinson, et al. 2020. Operationalising digital soil mapping – Lessons from Australia. Geoderma Reg. 23: e00335. doi: https://doi.org/10.1016/j.geodrs.2020.e00335.

Kuang, B., and A.M. Mouazen. 2012. Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. Eur. J. Soil Sci. 63(3): 421–429. doi: https://doi.org/10.1111/j.1365-2389.2012.01456.x.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, et al. 2022. caret: Classification and Regression Training.

Lagacherie, P., D. Arrouays, H. Bourennane, C. Gomez, and L. Nkuba-Kasanda. 2020. Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. Geoderma 375: 114503. doi: https://doi.org/10.1016/j.geoderma.2020.114503.

Lai, Y.-Q., H.-L. Wang, and X.-L. Sun. 2021. A comparison of importance of modelling method and sample size for mapping soil organic matter in Guangdong, China. Ecol. Indic. 126: 107618. doi: https://doi.org/10.1016/j.ecolind.2021.107618.

Lawrence, P.G., W. Roper, T.F. Morris, and K. Guillard. 2020. Guiding soil sampling strategies using classical and spatial statistics: A review. Agron. J. 112(1): 493–510. doi: https://doi.org/10.1002/agj2.20048.

Levi, M.R. 2017. Modified Centroid for Estimating Sand, Silt, and Clay from Soil Texture Class. Soil Sci. Soc. Am. J. 81(3): 578–588. doi: 10.2136/sssaj2016.09.0301.

Levi, M.R., and C. Rasmussen. 2014. Covariate selection with iterative principal component analysis for predicting physical soil properties. Geoderma 219–220: 46–57. doi: 10.1016/j.geoderma.2013.12.013.

Li, X., G.W. McCarty, L. Du, and S. Lee. 2020. Use of Topographic Models for Mapping Soil Properties and Processes. Soil Syst. 4(2). doi: 10.3390/soilsystems4020032.

Liang, P., C. Qin, A. Zhu, Z. Hou, N. Fan, et al. 2020. A case-based method of selecting covariates for digital soil mapping. J. Integr. Agric. 19(8): 2127–2136. doi: 10.1016/S2095-3119(19)62857-1.

Liao, K., S. Xu, J. Wu, and Q. Zhu. 2013. Spatial estimation of surface soil texture using remote sensing data. Soil Sci. Plant Nutr. 59(4): 488–500.

Loiseau, T., D. Arrouays, A.C. Richer-de-Forges, P. Lagacherie, C. Ducommun, et al. 2021. Density of soil observations in digital soil mapping: A study in the Mayenne region, France. Geoderma Reg. 24: e00358. doi: https://doi.org/10.1016/j.geodrs.2021.e00358.

Long, J., Y. Liu, S. Xing, L. Qiu, Q. Huang, et al. 2018. Effects of sampling density on interpolation accuracy for farmland soil organic matter concentration in a large region of complex topography. Ecol. Indic. 93: 562–571. doi: https://doi.org/10.1016/j.ecolind.2018.05.044.

Lopez, V.M.D. 2020. Natural Resources Conservation Service (NRCS). Salem Press Encycl. Sci. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=ers&AN=89474787&site=eds-live&custid=uga1.

Lu, S., B. Liu, Y. Hu, S. Fu, Q. Cao, et al. 2020. Soil erosion topographic factor (LS): Accuracy calculated from different data sources. CATENA 187: 104334. doi: 10.1016/j.catena.2019.104334.

Ma, Y., B. Minasny, B.P. Malone, and A.B. Mcbratney. 2019. Pedology and digital soil mapping (DSM). Eur. J. Soil Sci. 70(2): 216–235. doi: 10.1111/ejss.12790.

Mallarino, A.P., and D. Wittry. 2000. How can we make intensive soil sampling and variablerate P and K fertilization cost-effective. The Integrated Crop Management Conf. Proceedings. Nov. p. 29–30

Malone, B.P., B. Minansy, and C. Brungard. 2019. Some methods to improve the utility of conditioned Latin hypercube sampling. PeerJ. doi: 10.7717/peerj.6451.

Malone, B.P., B. Minasny, and A.B. McBratney. 2017. Using R for Digital Soil Mapping. Springer International Publishing, Cham.

Markewich, H.W., M.J. Pavich, and G.R. Buell. 1990. Contrasting soils and landscapes of the Piedmont and Coastal Plain, eastern United States. Geomorphology 3(3): 417–447. doi: https://doi.org/10.1016/0169-555X(90)90015-I.

Maxwell, A.E., T.A. Warner, and F. Fang. 2018. Implementation of machine-learning classification in remote sensing: An applied review. Int. J. Remote Sens. 39(9): 2784–2817.

McBratney, A., D.J. Field, and A. Koch. 2014. The dimensions of soil security. Geoderma 213: 203–213.

McBratney, A.B., M.L. Mendonça Santos, and B. Minasny. 2003a. On digital soil mapping. Geoderma. doi: 10.1016/S0016-7061(03)00223-4.

McBratney, A.B., M.L. Mendonça Santos, and B. Minasny. 2003b. On digital soil mapping. Geoderma. doi: 10.1016/S0016-7061(03)00223-4.

McBratney, Alex.B., B. Minasny, and U. Stockmann, editors. 2018. Pedometrics. Springer International Publishing, Cham.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of State Calculations by Fast Computing Machines. J. Chem. Phys. 21(6): 1087–1092. doi: 10.1063/1.1699114.

Minasny, B. 2013. Digital Mapping of Soil Carbon. Adv. Agron. 118: 1. doi: 10.1016/B978-0-12-405942-9.00001-3.

Minasny, B., and A.B. McBratney. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32(9): 1378–1388. doi: 10.1016/j.cageo.2005.12.009.

Mulla, D.J., and A.B. McBratney. 2001. Soil spatial variability. Soil Physics Companion

Najmadeen, H., O. Mohammad, and H. Mohamed-Amin. 2010. Effects of Soil Texture on Chemical Compositions, Microbial Populations and Carbon Mineralization in Soil. J Exp Biol 6(1): 59–64.

Natural Resources Conservation Service. Descr. SSURGO Database NRCS Soils. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_053627.

Nutrient Supply to Floodplains | EARTH 111: Water: Science and Society. https://www.e-education.psu.edu/earth111/node/823 (accessed 15 March 2022).

Olorunfemi, I., J. Fasinmirin, and A. Ojo. 2016. Modeling cation exchange capacity and soil water holding capacity from basic soil properties. EURASIAN J. SOIL Sci. EJSS 5(4): 266. doi: 10.18393/ejss.2016.4.266-274.

Omuto, C., F. Nachtergaele, and R.V. Rojas. 2012. State of the Art Report on Global and Regional Soil Information: Where are we? Where to go? : 81.

Padarian, J., B. Minasny, and A.B. McBratney. 2020. Machine learning and soil sciences: a review aided by machine learning tools. SOIL 6(1): 35–52. doi: 10.5194/soil-6-35-2020.

Pahlavan-Rad, M.R., and A. Akbarimoghaddam. 2018. Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). CATENA 160: 275–281. doi: 10.1016/J.CATENA.2017.10.002.

Peterson, A.M., W.H. Helgason, and A.M. Ireson. 2019. How Spatial Patterns of Soil Moisture Dynamics Can Explain Field-Scale Soil Moisture Variability: Observations From a Sodic Landscape. Water Resour. Res. 55(5): 4410–4426. doi: 10.1029/2018WR023329.

Piikki, K., J. Wetterlind, M. Söderström, and B. Stenberg. 2021. Perspectives on validation in digital soil mapping of continuous attributes—A review. Soil Use Manag. 37(1): 7–21. doi: 10.1111/sum.12694.

Ponizovsky, A.A., S.M. Chudinova, and Y.A. Pachepsky. 1999. Performance of TDR calibration models as affected by soil texture. J. Hydrol. 218(1): 35–43. doi: https://doi.org/10.1016/S0022-1694(99)00017-7.

Pusch, M., A.L.G. Oliveira, J.V. Fontenelli, and L.R. do Amaral. 2021. SOIL PROPERTIES MAPPING USING PROXIMAL AND REMOTE SENSING AS COVARIATE. Eng. Agríc. 41(6): 634–642. doi: 10.1590/1809-4430-eng.agric.v41n6p634-642/2021.

Quinn, P., K. Beven, P. Chevallier, and O. Planchon. 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrol. Process. 5(1): 59–79. doi: 10.1002/hyp.3360050106.

R Studio Team. 2015. R Studio.

Redlands, C.E.S.R.I. 2011. ArcGIS Desktop: Release 10.

Refaeilzadeh, P., L. Tang, and H. Liu. 2009. Cross-Validation. Encyclopedia of Database Systems

Robinson, A.C., U. Demšar, A.B. Moore, A. Buckley, B. Jiang, et al. 2017. Geospatial big data and cartography: research challenges and opportunities for making maps that matter. Int. J. Cartogr. doi: 10.1080/23729333.2016.1278151.

Rodriguez-Galiano, V., M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71: 804–818.

Rossiter, D.G., J. Liu, S. Carlisle, and A.-X. Zhu. 2015. Can citizen science assist digital soil mapping? Geoderma 259–260: 71–80. doi: https://doi.org/10.1016/j.geoderma.2015.05.006.

Roudier, P. 2011. clhs: a R package for conditioned Latin hypercube sampling.

Rounsevell, M.D.A., F. Ewert, I. Reginster, R. Leemans, and T.R. Carter. 2005. Future scenarios of European agricultural land use: II. Projecting changes in cropland and grassland. Agric. Ecosyst. Environ. 107(2): 117–135. doi: https://doi.org/10.1016/j.agee.2004.12.002.

Rowlandson, T.L., A.A. Berg, P.R. Bullock, E.R. Ojo, H. McNairn, et al. 2013. Evaluation of several calibration procedures for a portable soil moisture sensor. J. Hydrol. 498: 335–344. doi: 10.1016/j.jhydrol.2013.05.021.

Saleh, A.M. 2018. Spatial Variability Mapping of Some Soil Properties in Jadwal Al_Amir Project/Babylon/Iraq. J. Indian Soc. Remote Sens. 46(9): 1481–1495. doi: 10.1007/s12524-018-0795-x.

Samuel-Rosa, A., G.B.M. Heuvelink, G.M. Vasques, and L.H.C. Anjos. 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243–244: 214–227. doi: 10.1016/j.geoderma.2014.12.017.

Sanchez, P.A., S. Ahamed, F. Carré, A.E. Hartemink, J. Hempel, et al. 2009a. Digital soil map of the world. Science. doi: 10.1126/science.1175084.

Sanchez, P.A., S. Ahamed, F. Carré, A.E. Hartemink, J. Hempel, et al. 2009b. Digital soil map of the world. Science. doi: 10.1126/science.1175084.

Scull, P., J. Franklin, O.A. Chadwick, and D. McArthur. 2003a. Predictive soil mapping: A review. Prog. Phys. Geogr. doi: 10.1191/0309133303pp366ra.

Scull, P., J. Franklin, O.A. Chadwick, and D. McArthur. 2003b. Predictive soil mapping: A review. Prog. Phys. Geogr. doi: 10.1191/0309133303pp366ra.

Seybold, C.A., R.B. Grossman, and T.G. Reinsch. 2005. Predicting Cation Exchange Capacity for Soil Survey Using Linear Models. Soil Sci. Soc. Am. J. 69(3): 856–863. doi: https://doi.org/10.2136/sssaj2004.0026.

Sheets, Keith R., and Jan MH Hendrickx. "Noninvasive soil water content measurement using electromagnetic induction." Water resources research 31.10 (1995): 2401-2409.

Shit, P.K., G.S. Bhunia, and R. Maiti. 2016. Spatial analysis of soil properties using GIS based geostatistics models. Model. Earth Syst. Environ. 2(2). doi: 10.1007/s40808-016-0160-4.

Singh, J., D.M. Heeren, D.R. Rudnick, W.E. Woldt, G. Bai, et al. 2020. Soil structure and texture effects on the precision of soil water content measurements with a capacitance-based electromagnetic sensor. Trans. ASABE 63(1): 141–152. doi: 10.13031/trans.13496.

Söderström, M., G. Sohlenius, L. Rodhe, and K. Piikki. 2016. Adaptation of regional digital soil mapping for precision agriculture. Precis. Agric. doi: 10.1007/s11119-016-9439-8.

Soil Survey Staff. 2018. Gridded Soil Survey Geographic (gSSURGO) Database for the Conterminous United States. U. S. Dep. Agric. Nat. Resour. Conserv. Serv.

Somarathna, S., B. Minasny, and B. Malone. 2017. More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. Soil Sci. Soc. Am. J. 81. doi: 10.2136/sssaj2016.11.0376.

Stevens, R.G. 1992. Soil resources: what is needed and how do we maintain these resources. Am. J. Potato Res. 69(11): 717.

Sun, X.-L., Q. Yang, H.-L. Wang, and Y.-J. Wu. 2019. Can regression determination, nugget-to-sill ratio and sampling spacing determine relative performance of regression kriging over ordinary kriging? CATENA 181: 104092. doi: https://doi.org/10.1016/j.catena.2019.104092.

Terribile, F., A. Coppola, G. Langella, M. Martina, and A. Basile. 2011. Potential and limitations of using soil mapping information to understand landscape hydrology. Hydrol. Earth Syst. Sci. doi: 10.5194/hess-15-3895-2011.

Thompson, J.A., J.C. Bell, and C.A. Butler. 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma 100(1): 67–89. doi: https://doi.org/10.1016/S0016-7061(00)00081-1.

Unesco, F.A.O. 1987. Soils of the World. Elsevier Science Publishers, Amsterdam.

Verberg, P.H., W. Soepboer, A. Veldkamp, R. Limpiada, V. Espaldon, et al. 2002. Modeling the Spatial Dynamics of Regional Land Use: The CLUE-S Model. Environ. Manage. 30(3): 391–405. doi: 10.1007/s00267-002-2630-x.

Villas-Boas, P.R., R.A. Romano, M.A. de Menezes Franco, E.C. Ferreira, E.J. Ferreira, et al. 2016. Laser-induced breakdown spectroscopy to determine soil texture: A fast analytical technique. Geoderma 263: 195–202. doi: 10.1016/J.GEODERMA.2015.09.018.

Wadoux, A.M.J.-C., D.J. Brus, and G.B.M. Heuvelink. 2019. Sampling design optimization for soil mapping with random forest. GEODERMA 355. doi: 10.1016/j.geoderma.2019.113913.

Webster, R., and M.A. Oliver. 1992. Sample adequately to estimate variograms of soil properties. J. Soil Sci. 43(1): 177–192. doi: https://doi.org/10.1111/j.1365-2389.1992.tb00128.x.

Wendroth, O., E.L. Ritchey, S. Nambuthiri, J.H. Grove, and R.C. Pearce. 2011. Spatial Variability of Soil Physical Properties. In: Gliński, J., Horabik, J., and Lipiec, J., editors, Encyclopedia of Agrophysics. Springer Netherlands, Dordrecht. p. 827–839

Were, K., D.T. Bui, Ø.B. Dick, and B.R. Singh. 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52: 394–403. doi: https://doi.org/10.1016/j.ecolind.2014.12.028.

Wilson, J.P., and J.C. Gallant, editors. 2000. Terrain analysis: principles and applications. Wiley, New York.

Wischmeier, W.H., and J.V. Mannering. 1969. Relation of Soil Properties to its Erodibility. Soil Sci. Soc. Am. J. doi: 10.2136/sssaj1969.03615995003300010035x.

Wollenhaupt, N. 1996. Sampling and testing for variable rate fertilization. Proceedings of the 1996 Information Agriculture Conference. P & PI Norcross, GA. p. 33–34

Zhang, Y., and A.E. Hartemink. 2021. Quantifying short-range variation of soil texture and total carbon of a 330-ha farm. Catena 201: 105200.

Zimmerman, D., C. Pavlik, A. Ruggles, and M.P. Armstrong. 1999. An experimental comparison of ordinary and universal kriging and inverse distance weighting. Math. Geol. 31(4): 375–390.

CHAPTER 2

OPTIMIZING SAMPLE SIZE FOR PREDICTING SOIL TEXTURE IN A FLOODPLAIN SOIL OF THE GEORGIA PIEDMONT, USA

[1]Platero, D., M. Levi, N. Gaur, D. Markewitz, D. Franklin To be submitted to Geoderma.

2.1 Abstract

Soil texture controls crucial processes such as water infiltration, flood extenuation, soil-plant relationships, and nutrient availability. High spatial resolution images as well as digital soil mapping are producing predictive soil texture maps with improved resolution. However, current soil databases lack the spatial resolution necessary for precision agricultural management at the field scale and determining the optimal sample size is often a challenge in crop fields. The objective of this work was to determine the optimal sample size for field soil collection to produce accurate predictions of soil texture at multiple depths to 1 m for a 40-acre crop field in a Georgia Piedmont floodplain. Soil samples from four depths were collected at 69 locations and analyzed for particle size distribution before being combined with environmental covariate data using random forest algorithms to predict sand, silt, and clay. Models were developed for 50 random iterations of varying sample sizes from 10 – 65 to compare the effect of sample size on model performance. We determined that 35-45 samples were sufficient for the study area based on the trends in R^2 and RMSE for sand and clay at 0 – 10 and 40 – 70 cm depths. Results from this study suggest a sample density of approximately 1 location per ac provides sufficient information for detailed soil texture mapping in the floodplain landscape we evaluated.

2.2. Introduction

Soil properties can express a high degree of spatial variation over short distances which can have pronounced impacts on management decisions and modeling efforts that require soil information (Franzen, 2018). The degree of spatial variability can also be further challenging to

represent for different soil depth increments (Peterson et al., 2019). Variability can result from inherent differences produced during soil development, management activities like tillage and subsequent erosion, and systematic errors from uneven use of fertilizers and manures (Franzen, 2018). Current soil maps like those available in the Soil Survey Geographic Database (SSURGO) and State Soil Geographic (STATSTGO; (Soil Survey Staff, 2018) are often of insufficient resolution for precision agricultural applications that employ very localized management for water and nutrients.

A variety of covariate data have been used to represent the processes and patterns of soil formation ranging from spectral reflectance to topography obtained from sensors on satellites, drones, tractors, and even hand-held sensors. For example, Pusch et al. (2021) tested whether remote and proximal sensing data could assist in soil property mapping in Brazil through geostatistical prediction. They chose two covariates often used to express soil variations—one obtained by remote sensing (a short-wave infrared band; SWIR2) and the other by proximal sensing (apparent soil electrical conductivity – ECa)—to compare them individually and together in a geostatistical interpolation method (kriging with external drift). They found that ECa was a more promising covariate than SWIR2 band from orbital imaging. Such proximal sensing can identify the soil short-range spatial variability. However, when the soil property variability is well explained by the sampling procedure, multivariate geostatistical methods may not improve the mapping accuracy (Pusch et al., 2021). Hummel et al. (1996) used two sensors to obtain site-specific data on factors affecting crop growth and yields, such as nutrient status, weed pressure, soil moisture status, landscape position, soil organic matter (SOM) content, soil acidity, and depth to a restrictive layer. The sensors used were a single-wavelength, soil catena-dependent sensor and a multiple-wavelength, catena independent sensor. Sensor-based estimates of soil

properties are usually less accurate than those obtained by lab analyses. Their study was located in the Midwestern, USA and they mention how the area has a large coverage of claypan soils. Subsequently, mapping of claypan depth has not been practical for production agriculture. In order to quantify deep variation over a large area, an automated, preferably non-invasive, measurement approach is required (Hummel et al., 1996). However, real-time sensing provides much more data with the same amount of effort, and these multiple points can be averaged to improve prediction accuracy if calibrated properly.

DSM provides an opportunity to increase the spatial resolution of existing largescale soil survey (e.g., SSURGO). One obvious limitation of the soil survey products is that most are polygon-based products that cannot capture spatial resolutions typically needed for site-specific applications. The polygons can be converted to raster datasets to facilitate more modeling approaches, but often still reflect the same polygonal units as the original maps as in the gridded Soil Survey Geographic dataset (gSSURGO). The need to have additional point data for improved maps from DSM is clearly recognized (Arrouays et al., 2020a), however, obtaining soil data is often limited by time and cost constraints (Arrouays et al., 2020b; Kidd et al., 2020; Lagacherie et al., 2020). Therefore, significant effort has been focused on the placement of sample locations to overcome the constraints of cost and time for sample collection (e.g., the conditioned latin hypercube sample design) (Roudier, 2011). However, the ideal number of points required for detailed maps is not always known at local mapping efforts where optimal sample size would be utilized. Having 2-3 soil samples per ha (Shannon, 2018; Mohamed, 1996) and resolution of 5 m or less for multispectral bands (Chenghai, 2018) and spatial resolution of lidar for precision agriculture applications should be between 2 to 5 square meters per pixel with a positional accuracy of within 2 meters (Moran, 2000).

The number of samples required for spatial modeling is dependent on the prediction models being used and the properties of interest. For example, kriging methods that require the development of a semi variogram often require a minimum of 50-100 samples for one area (Webster and Oliver, 1992; Hengl et al., 2007a; Levi and Rasmussen, 2014) whereas other interpolation techniques such as inverse distance weighting can be accomplished with fewer points (Zimmerman et al., 1999; Harper, 2006). Machine learning has been applied extensively in the DSM literature with some of the most common techniques being RF, support vector machines, and other decision tree-based approaches (Rodriguez-Galiano et al., 2015; Brungard et al., 2015; Heung et al., 2016; Maxwell et al., 2018). When a design-based approach is used to estimate whole-field or within-stratum spatial means and variances, determining the appropriate sample size is important for achieving the desired precision and accuracy (Lawrence et al., 2020). The variability in soil moisture is also a good indicator of variability in soil texture since it is controlled by the latter (Gaur and Mohanty, 2013, 2016). Duffera et al. (2007) conducted research in the Coastal Plain region of North Carolina using 60 coil cores with a ~ 1 m depth with five depth increments. Their study area was 12 ha in North Carolina. Here they split the samples into two, first was particle size distribution and second was bulk density, saturated hydraulic conductivity, and porosity. They concluded that using just particle size density was the best option for spatial variability for management zones and this approach could cover 62% of the area. The grids in their study resulted in three soil map units being sampled in approximate proportion to their areal extent in the field: Goldsboro, 6.9 ha, n=40; Lynchburg, 3.0 ha, n= 14, and Norfolk, 0.95 ha, n=6 (Duffera et al., 2007). For precision agriculture applications, the convention is often to grid sample with sample densities of previous studies and have ranged from 1-4 samples per ac (Ferguson and Hergert, 2009; Kerry, 2010). Grid sampling can be

random, random cluster, or systematic (Wollenhaupt, 1996). Franzen (2018) used grid sampling and recommends one sample per hectare in places like Iowa, Illinois, and Indiana. There is high variability in these fields due to fertilizer buildup resulting in high soil test values with similar recommendations. Because of the uniformity of recommendation, a 2.5-acre grid is acceptable in these types of fields. Franzen (2018) mentions that if there is a high variability in the recommendation, then a higher sampling density may be required to create an accurate map.

A lot of effort has been spent on developing algorithms and techniques to estimate the spatial distribution of soil texture given the importance of understanding its spatial variability, but most studies conclude that results can be improved by altering the sample size used for algorithm development. For example, Liao et al. (2013) indicate that their kriged and multiple stepwise regression techniques to estimate spatially distributed soil texture in the PingduShandong Province of China suffered as a result of inadequate sample size for that landscape. In another study over the Southern Great Plains in the U.S., Chang et al., (2003) identified the distribution of samples and sample size affected their neural network algorithm performance for determining soil texture. Zhang and Hartemink (2021) recommend increased sampling density to improve spatial estimates of soil properties. Their study area was 330 ha located in southcentral Wisconsin, USA. They measured soil properties at short range variations which affected spatial structure variograms. Many studies recognize the influence of sample size on spatial prediction models of soil properties, however, an adequate sample size density for these efforts lacks consensus.

The purpose of this study is to determine the optimal sample size for representing the spatial variability in soil texture and distinguish the spatial variability at different depths between 0-100 cm in a typical floodplain of the Georgia Piedmont that is managed for agricultural use.

2.3. Materials and Methods

2.3.1 Study area

The study area for this research is a 40-acre crop field located at the University of Georgia Iron Horse Farm 14 miles south of Watkinsville, GA (Fig. 1). It sits on a floodplain adjacent to the Oconee River in the Piedmont region of Georgia. Near the Oconee River section of the field is ~ 134 m and elevation slowly climbs as you go west to 145 m. Geologically, the Piedmont is primarily comprised of metamorphic rocks, but includes granitic intrusive bodies. These rocks have been weathered to thick saprolite (2-20 m) over much of the region (Frazier, 2006). The soils of the region have been subject to intensive weathering that dissolved or altered almost all primary minerals and left behind a residue of clays containing aluminum and iron oxides (Lester and Allen, 1950). Upland soils in the region are classified as Ultisols with red, clay-rich subsoils with a low base saturation (Markewich et al., 1990). Alluvial soils in the region represent Entisols and Inceptisols with varying thicknesses of sediment from both natural and anthropogenic drivers. For example, the Piedmont region has many regions of legacy sediment resulting from extensive erosion that occurred over the last 200 years following European settlement (Donovan et al., 2015; "Georgia Soil Survey 136 - Southern Piedmont | NRCS Georgia,").

The research area (Figure 3) is mapped as having Chewacla, Wehadkee, and Wickham soil series (Soil Survey Staff, 2018). Chewacla soils form from alluvial sediments in floodplain and classified as Fine-loamy, mixed, active, thermic Fluvaquentic Dystrudepts with a drainage class of somewhat poorly drained and flooding frequency and duration that is rare for very brief to lengthy periods. The Wehadkee soils are classified as Fine-loamy, mixed, active, nonacid, thermic Fluvaquentic Endoaquepts consisting of very deep, poorly drained, and very poorly

drained soils on flood plains along streams that drain from the mountains and piedmont. They are formed in loamy sediments. The Wickham series is classified as Fine-loamy, mixed, semiactive, thermic Typic Hapludults consisting of very deep, well drained, moderately permeable soils on stream terraces in the Piedmont and Coastal Plain. Argillic horizon is present in the Wickham.

The field is situated at 139 meters above sea level and annual precipitation averages ~1100 mm (43 inches) (Suleiman and Hoogenboom, 2007) and mean annual air temperature is 20-22ºC ("Georgia Weather - Automated Environmental Monitoring Network Page,"). Elevation is highest in the southwest (Fig.3) corner of the study area, which contributes runoff into the lower elevation on the northeast portion of the field closer to the floodplain and the influence of the Oconee River. It has historically (1993-2016) been used for hay/pasture as seen on Google Earth Pro images from USGS, USDA Farm Service Agency, and Landsat. Tile drains were installed in 2015 and are ~80 cm beneath the surface. In June of 2020, soybeans (Glycine max) were planted after site preparation that included disking and rototilling which allowed for a near bare soil condition for the collection of surface reflectance.

2.3.2 Topographic Covariates

The internal flow of water in landscapes affects soil texture transformation (Franzen, 2018). A digital elevation model (DEM) derived from Lidar obtained from NOAA (National Oceanic and Atmospheric Administration) (Conrad et al., 2015) (2012 GA DNR Lidar) was used to develop topographic indices using the module of compound analysis in the open-source System for Automated Geoscientific Analyses (SAGA) software (Conrad et al., 2015) to represent the study area. The spatial resolution of the LiDAR was 1.2 m and all topographic covariates also had 1.2 m resolution. Terrain analysis utilize geomorphometric calculations to

develop representations of slope, aspect, analytical hill shading, and flow paths (Böhner and Selige, 2006). Topographic covariates used in this study included length slope factor, flow path length, multi scale topographic position index, topographic position index, saga wetness index, digital elevation model, multiresolution index of ridgetop flatness, and multiresolution index of valley bottom flatness. A more detailed description of the covariates is provided in Table 2.

Three topographic covariates, namely flow path length, saga wetness index and topographic position index were used in the sample design of this study. Flow path length allows for the calculation of several terrain indices from a digital elevation model (Böhner et al., 2001; Böhner and Selige, 2006). It calculates average flow path starting at particular point locations. This point location points runoff processes and calculates the runoff origin, meaning it calculates upstream and downstream flow (Quinn et al., 1991). The SAGA wetness index is based on a modified catchment area calculation ('Modified Catchment Area'), which allows for the flow of water from one cell to multiple adjacent cells instead of a single cell like a deterministic flow algorithm does (e.g., D-8 used in a standard topographic wetness index calculation). As a result, it provides a more realistic, higher potential soil moisture compared to the standard topographic wetness index calculation (Böhner et al., 2001). The topographic position index is identical to the difference to the mean calculation (residual analysis) proposed by Wilson & Gallant (2000). Another significant factor is the LS slope factor. The crop field under consideration has tillage, rotation, and erosion happening which makes the LS factor a beneficial covariate. According to Lu et al. (2020), the availability and precision of topographic data controls the reliability of the calculated slope length and slope gradient.

2.3.2.1 Spectral Covariates

In addition to the topographic covariates derived from Lidar, additional covariates were derived from drone images that were obtained three days after planting soybeans to represent patterns of soil variability in surface soils. These covariates included visible multi-spectral images collected on June 12, 2020. Red and green spectral imagery was collected with a pixel with a dimension of 0.01118 m per side. Complications with the blue band precluded its use for this study. The field condition was predominantly bare soil, however, soybeans had emerged and were approximately 10 cm high. The high-resolution imagery was resampled to 1.2 m with the cubic convolution method in ArcGIS to make the resolution consistent with the topographic covariates. Red and green bands were combined using a principal component analysis in ArcMap on a set of the green and red raster bands and generated a single multiband raster as output (Redlands, 2011).

2.3.2.2 EM-31

Apparent electrical conductivity of the soil was obtained using an EM-31. Soil apparent electrical conductivity (ECa) measured by electromagnetic induction (EMI) has been widely used to interpret soil spatial variability. The EM-31 maps geological variations, groundwater contaminants or any subsurface feature associated with changes in the ground conductivity using an electromagnetic inductive technique that makes the measurements without electrodes or ground contact (Dadfar et al., 2011). The EM-31 data was collected on the farm with a fiberglass cart so there was no metal to interfere with readings. The instrument is ~ 3 meters long and was pulled by a UTV. The EM-31 was used in a single orientation (north-south) for this project such that the depth of measurement was up to three meters beneath the ground surface. We collected

measurements on March 23, 2021, with south to north transects to capture different soil moisture conditions. Field conditions for the day was dry in most areas and was at field capacity. The data collected from the EM-31 was used as a covariate for the prediction soil texture mapping of the study area. Data from the EM-31 was taken every second with a total of 29,268 data points. Inverse distance weighting was used in ArcMap to create the raster.

2.3.3 Sample Design

Soil samples were collected using a combination of two sampling designs. A conditioned Latin Hypercube sample design (cLHS;(Minasny and McBratney, 2006)) algorithm was used to identify 50 locations within the study area (Phase 1) and another set of samples was collected from 20 separate locations using a transect approach spanning the entire field (Phase 2). The sampled points are shown in Fig. 1. The cLHS was implemented using the cLHS package in R (Godinho Silva et al., 2015; R Studio Team, 2015) to identify 50 soil sample locations representing the spatial variability of microtopography and anticipated soil variability using three environmental covariate layers (SAGA wetness index, flow path length, topographic position index; Fig. 4 for Phase 1 sampling. The cLHS algorithm has been used for planning field sampling surveys in order to understand the spatial behavior of natural phenomena such as soils by capturing the variance in a matrix of covariates to improve quantitative prediction of special variance. The three covariates were chosen based on their ability to capture spatial patterns of slope and hydrological properties (Malone et al., 2019). The sample sets are optimized using simulated annealing (Metropolis et al., 1953) through a set of k iterations to represent a Latin hypercube. Simulated annealing mimics the controlled cooling process used to reach a global optimum. At each iteration *k* of the simulated annealing, changes are made to the sampling

scheme S(k−1) to form a new candidate sampling scheme S(k). An objective function, detailed by Minasny and McBratney (2006), assesses how well S(k) represents a Latin hypercube. The resulting objective function value, obj(k), is used to compute $M_{obj}(k)$:

$$M_{obj}(k) = e(-\frac{\Delta_{obj}(k)}{T(k)}) \quad (1)$$

Where Δobj(k) = obj(k) – obj(k-1), The variation in the objective function, and T(k) is the current temperature at iteration k, which is decreased by a cooling factor δ every p iterations (Roudier et al., 2012).

2.3.5 Soil Sampling and Spatial Predictions

At each sampling location, soils were collected from 0-10, 10-40, 40-70, and 70-100 cm depths on June 15 and 16, 2020 using a hand auger. Samples at each location were composited in a 5-gallon plastic bucket by depth, thoroughly mixed by hand, and a ~500-gram sample was retained to represent the depth increment. Soils were air-dried and sieved through a 2-mm screen (Fig. 2) prior to analysis with a Beckman Coulter LS 13 320 Particle Size Analyzer (LPSA) to determine particle size distribution and percentages of sand, silt, and clay. The air-dried soil samples were split using a sample splitter (SP-3, Lewis Center, OH) to obtain a representative subsample and reduce bias for laboratory replications. Subsamples of 0.5-gram were weighed into 15-ml centrifuge tubes and 5 ml of sodium hexametaphosphate (5% solution) was added before shaking for 15 hours at 120 oscillations per minute to disperse the soil particles. After shaking, the soil solution in the centrifuge tube was transferred to a 13-ml test tube using deionized water and filled nearly to the top. Samples were placed in the LPSA carousel for particle size analysis. Two 60-second analyses were averaged for each laboratory replicate with

the Polarization Intensity Differential Scattering (PIDS) functionality enabled and sonication during analysis (i.e., analytical replicates). Once analysis was completed, the data was saved and exported. Two laboratory replicates were analyzed for quality control and a third replicate was analyzed for any sample with > 5% differences in either clay or sand from the first two replications.

Measured sand, silt, and clay data determined for each sampling location was combined with environmental covariate data to evaluate relationships between covariates and soil texture at multiple depths. Pearson correlations were evaluated between covariates and measured point data to determine Pearson's correlation coefficient using a Row-wise method in JMP (Figs. 6 & 7) (JMP®, 2021). Soil-landscape models for spatial predictions were developed with a random forest framework (Gessler et al., 1995) in R (R Studio Team, 2015) using the randomForest (R Core Team, 2017) and Caret packages (Max, 2008). Random forest models consist of many individual decision trees that act as an ensemble (Breiman, 2001). The ensemble uses multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. A forest that is a set of randomized decision-making trees is built and trained based on a bootstrap approach. The trees in the ensemble are built on the basis of the principle of recursive partition (Amirian-Chakan et al., 2019). Strobl et al., (2009) states the feature space is recursively split into regions containing observations with similar response values. The predictions of individual trees are then averaged to give a single prediction (Amirian-Chakan et al., 2019). Every decision tree in the forest is trained on a subset of the dataset called the bootstrapped dataset. For this model, an 80/20 split was used, meaning 80 percent was used for training and 20 percent for prediction. Following (Kuhn et al., 2022), k-fold cross validation was applied using the caret package to test the model performance. In k-fold

cross validation, the data is first divided into k near-equally sized folds. Then k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining k − 1 folds are used for learning (Refaeilzadeh et al., 2009; Piikki et al., 2021). Cross validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In typical cross validation, the training and validation sets must cross over in successive rounds such that each data point has a chance of being validated against (Refaeilzadeh et al., 2009). Validation provides information on how well a particular model performed in practice and is considered a crucial task associated with any DSM project (Piikki et al., 2021).

2.3.5.1 Evaluating impact of sample size

The performance of the model was computed from the validation samples by calculating the correlation between the observed and estimated values based on the coefficient of determination (R^2) and root mean square error (RMSE). For each property and depth combination (sand, silt, and clay for 4 depths), 50 iterations were run by randomly selecting different sub-samples from the dataset for testing the impact of sample sizes on soil texture prediction. The sample sizes represented all sequences of 5 between 10-65. Cross validation was used to validate the model and R^2 and RMSE were calculated for each iteration. R^2 of each model iteration was computed using a linear model between observed and predicted values with the lm function in R (Everitt, 1992; R Core Team, 2017). The R^2 is a measure that provides information about the goodness of fit of a model. In the context of regression, it is a statistical measure of how well the regression line approximates the actual data. The equations below are as follows (Johnston et al., 2001):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}[Z(X_i) - \hat{Z}(x_i)]^2} \quad (2)$$

$\hat{Z}(x_i)$ is the predicted value, ($Z(x_i)$ is the observed value, N is the number of values and for location i.

$$R^2 = 1 - \frac{Sum\ squared\ regression\ (SSR)}{Total\ sum\ of\ squares\ (SST)} \quad (3)$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum y_i - \hat{y})^2}$$

Residual= actual y value – predicted y value

Σ= sum

$$r_i = y_i - \hat{y}_i$$

The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared. As a percentage, it will take values between 0 and 1 (Anderson-Sprecher, 1994).

2.4. Results

2.4.1. Details of measured soil properties

Sand, silt, and clay percentages of the measured samples varied considerably for each depth (Table 1 and Fig. 5). For presentation of results and discussion, 0-10 and 10-40 cm depths are considered shallow and 40-70 and 70-100 cm depths are considered deep. Additionally, the strongest emphasis is on sand and clay for 0-10 and 40-70 cm depths. For shallow depths, clay, clay loam, and loam dominated the soil texture while deeper depths ranged from sandy loam, loam, clay loam, and more clay dominated than surface soil (Fig. 5). Clay and sand had the

widest range of values across all depths with clay having an average 48.4% range and sand 51.1%. Silt had the lowest range with an average of 37.05% across all four depths.

Clay percentages increased with depth starting with a mean of 28.3 % at 0-10 cm and 46.0 % at 70-100 cm. At the 0 – 10 cm depth, clay was moderately skewed (0.80) and approximately symmetric for other depths. Clay was the only soil property that had a negative skewness at -0.2 and -0.3%. Kurtosis was at 4.1 at 0-10 cm and 1.9 at 70-100 cm depth, making it a platykurtic shape for flatter distribution. The range values were high which means there was greater variability. It should be noted that high-range values mean a higher standard deviation (SD) in clay at all depths. Clay had a larger STD reflecting a wide range of 46-53% values indicating notable variability in the field.

The average sand percentage was between 30.7% at 0-10 cm and 23.6% at a depth of 70-100 cm. The percentage did not change as much as clay. The skewness increased with depth from 0.40 at 0-10 cm to 1.2 at the 70-100 cm depth. Kurtosis for sand for all depths was >3 indicating a more pronounced peak than clay (leptokurtic shape). Sand had a range starting at 43.2% at 0-10 cm and slowly increased to 57.9% at 70-100 cm. Sand had higher range values than clay and silt and the range values increased with depth. This indicated that sand had the greatest variability of all measured soil properties. At the shallow depths (0-10 & 10-40 cm), STD was less than that of deeper depths (40-70 & 70-100 cm). Starting at a STD of 9.1% and ending with 13.3% reflecting the similar trends in the range.

Silt had a similar distribution across all depths with an average of 41% at 0-10 cm and 30.4% at 70-100 cm. The range was at 32.2% at the surface and slowly increased to 41% at 70-100 cm. Of the three soil properties, silt had the lowest variability in the range for all depths. There was also a trend of SD increasing with depth from 6.7% at 0-10 cm to 9.5% at 70-100 cm.

The data tended to slowly spread away from the mean values with increasing depth. Kurtosis had an average of 3% for silt making it mesokurtic shaped and less varied kurtosis than clay and sand. Skewness averaged 0.04% indicating a slight positive skew.

.

2.4.2. Relationship of environmental covariates and soil properties

The ten covariates used for modeling in this study showed diverse degrees of correlation with soil properties as seen in Figs. 6 and 7. At 0-10 cm silt had the strongest correlations with covariates whereas clay had the strongest correlations with covariates at the 40-70 cm depth. In both scenarios, the elevation and landform had the strongest relationships with measured soil texture. Sand had weak relationships between covariates for both depths. At 0-10 cm, clay had weak positive correlations with TPI_Landform, DEM, LSF, and FPL. The environmental covariates that were negatively correlated, especially TPI_Landform, showed some soil surface relationships between shallow depths and topography where the high values of TPI_Landform were where the surface silt was high in percentage. High LSF values at ~ 5 and FPL values at ~150-100 were also where clay percentage was high. The field has an area where the soil is saturated most of the time in the floodplain near the Oconee River with less clay and more silt (Fig. 1).

2.4.3. RF model performance

Violin plots showing R^2 and RMSE percentages for different sample sizes used for prediction are shown in Figures 9-12. Model accuracy was performed using data from the 69 sample points in the RF model. For this study, the depths of 0-10 and 40-70 cm are discussed in

detail to reflect the variability of processes in the soil profile. The 0-10 cm depth is the interface between soil and the atmosphere and is important for plant establishment whereas the 40-70 cm depth better reflects soil texture controls on subsurface moisture dynamics and water table fluctuations. Figs. 8, 9, 10, and 11 show the performance of RF models for each sample size by representing the distribution of R^2 and RMSE for clay and sand percentages. A common trend in all predictions was that the R^2 increases while RMSE decreases with an increase in sample size. The uncertainty in predictions also decreases with an increase in sample size.

2.4.3.1. Clay Performance

At 0-10 cm, as expected, the lower sample size had the lowest R^2 and highest RMSE (Figs. 8 and 9). For the sample size of 10, R^2 was 0.2 and RMSE of 5%. As the sample size increased, the variability of model performance from the 50 random iterations began to decrease. At a sample size of 35, the R^2 was 0.76 and a RMSE of 3.3%. At 65 samples, the R^2 was 0.90 with an RMSE of 2.3%. For 40-70 cm plots, the R^2 was at 0.60 and RMSE at 13%. There were outliers for every distribution of model performance indicating that some soil data didn't perform well with covariates. Sample size 35 had an R^2 of 0.82 and RMSE of 6%. Sample sizes 40-60 had a plateau of R^2 at 0.88 while sample size 65 has an R^2 of 0.92 and RMSE of 3.1%. High LSF values at ~ 5 and FPL values at ~150-100 were also where clay percentage was high.

2.4.4. Sand Performance

Sand at 0-10 cm had a wide range of distribution at 10,15, and 20 samples of R^2. The RMSE was at 8% at a sample size of 10 and decreased as sample size increased. Sample sizes 15 and 20 had the most outliers in the plots in R^2 and RMSE. At a sample size of 35, the R^2 was 0.80 an RMSE at 7.8%. At 65 samples, the R^2 was 0.93 and RMSE was at 3.1%. The lower

performance of the model for sand as compared to clay especially for small sample sizes was due to the larger range of the percentage of sand in the field. A smaller sample size may not represent the range in its entirety and the decision trees that quantify the relationships between the covariates and sand may vary with percentage of sand.

2.4.5. Variable Importance for RF models

Variable importance differed by modeled soil property and depth as indicated by the increase in (MSE) Mean Decrease Accuracy derived from RF models Figs. 12-15. The results are presented for three sample sizes of 20, 35 and 69 that represent sample sizes where the RF model performance was poor (20), 35 where the RF model performance showed improvement for all three textural class predictions and 69 which was the maximum possible sample size. Clay MSE in 0-10 cm of covariate prediction correlated well with the statistical summary as seen in Table 1. FPL was the highest predictor in all samples of 20, 35, and 69. Aspect, FPL, SWI, and DEM were in the top four of covariate prediction power in 0-10 cm. The predictive power of FPL for Clay in 40-70 cm was lowest while DEM was the highest in all three sample sizes of 20, 35, and 69. The covariates that had the lowest predicting power at 0-10 cm were strongest predictors in the 40-70 cm depths like MRVBF, MRRTF, and EM.

The dominant covariates also changed with sample size for the same soil property and depth which indicates that sample size influences the sensitivity of different covariates to soil property prediction. At 0-10 cm, FPL is the strongest predictor but at 69 samples, decreases. This follows a trend where the strongest predictors at 20 samples decrease as sample sizes get larger. For example, FPL and PCA were the strongest predictors for 20 samples at the 0-10 cm depth. However, at 35 samples, PCA and SWI were the strongest predictors while at 69 samples, DEM

and PCA were the strongest predictors. At 40-70 cm depth, 20 samples had FPL as the strongest predictor covariate. At 40-70 cm, FPL and TPI_Landform followed a trend where it was the strongest predictor at 20 samples but at 69 samples, they are at the bottom. Another trend was the weak predictors (i.e., MRRTF and PCA) increased as samples sizes increased. At 35 samples EM was the strongest predictor while at 69 samples, it is the weakest predictor. Finally, 69 samples had MRRTF as the strongest predictor.

2.4.6. Spatial variation of properties

Figs. 12 & 13 show maps of spatial distribution of clay for 0-10 cm and 40-70 cm when using sample sizes of 20, 35 and 69 for prediction along with the relative importance of different covariates used in the models for different sample sizes. The covariates with the strongest predictive power did not vary as much with sample sizes but changed considerably with depth. At the 0-10 cm depth, the FPL covariate is the most important predictor of all three sample sizes for clay. At the 40-70 cm depth, elevation had the highest predictive power for clay. Elevation with < 138 m (Fig.3) has low clay percentages of 20-30% shown in Fig.12. For the 0-10 cm depth, model performance increased with sample size from an R^2 of 0.51 and RMSE of 7.21 % for 20 samples to an R^2 of 0.69 and RMSE of 6.14% for 35 samples. The model using all 69 samples performed very well with an R^2 of 0.94 and RMSE of 1.19 %.

Prediction maps of clay at 40-70 cm showed more spatial variability within the field than the 0-10 cm predictions. Areas of the field with low ECa tended to have higher clay percentages. Observing the maps created with 20, 35, and 69 samples, there is an area in the lowest portion of the field with low clay. For 40-70 cm depth, DEM is the strongest covariate predictor for all three sample sizes. TPI_Landform is also high for 20 and 69 sample sizes and is reflected in the

prediction maps by the marked boundaries between high and low estimates. SWI is the next strongest predictor for 35 samples. The spatial patterns of 35 and 69 sample sizes are nearly identical. As the sample size increases, the percentage of clay scatters most of the northwest corner of the field which shows a 40% clay pattern. The map created with 20 samples showed large areas of low clay percentage visible on the prediction maps, and these same patterns are found on the TPI_Landform covariate. At 40-70 cm depth, model performance increased with sample size from an R^2 of 0.51% and RMSE of 7.53% for 20 samples. At 35 samples, the R^2 is 0.80 and RMSE of 5.94%. The model using 69 samples performed well with an R^2 of 0.88 and RMSE at 3.75%.

Figs.14-15 a-c show maps of spatial distribution of sand for 0-10 and 40-70 cm when using sample sizes of 20, 35 and 69 for prediction. Figs. 14-15 d-f show the relative importance of different covariates used in the models for different sample sizes. In contrast to clay, the covariates with the strongest predictive power for sand vary with both sample sizes and depth. Observing DEM, sand tends to be high in percentage as elevation increases as seen in the prediction map and DEM covariate. At 0-10 cm depth with sample sizes of 20, 35, and 69, while the range of percentage of sand is large within the field, there is little spatial variation. The greatest amounts of sand are found in the northern section of the field with percentages ranging from 33 - 45%. Unlike clay, sand percentages were lower in the higher elevation areas and similarly low in the lowest landscape position in the study area. Topographically, sand had high percentages in the washes in the northern section. At 20 samples, FPL and surface reflectance (PCA) were the best predictors as covariates. At a sample size of 35, FPL was the strongest predictor followed by SWI. TPL_Landform was the lowest predictor covariate. At the 40-70 cm depth, there was more variability between the sample sizes. For a sample size of 20, FPL,

TPI_Landform, and DEM were the most important covariates, but these same covariates were less important for the other sample sizes. One of the reasons why LSF is a poor predictor is because it calculates the effect of slope length on erosion and is the ratio of soil loss from field slope length. Therefore, FSL is usually a better predictor for surface soils rather than deeper soils. SWI was generally not an important variable for predicting sand in this study area. At 35 samples, ECa was the highest predictor covariate and TPI_Landform being the lowest. Sand percentage in spatial patterns decreased from 20 samples. At 69 samples, with a R^2 of 0.92 and RMSE of 2.8%, majority of sand at 35-40% decreased significant only being found in areas of the floodplain and washes. Much of the field remains at 10-25% sand mainly in higher-altitude areas. MRRTF and PCA were the two most important predictors for sand at 69 samples at 40-70 cm. EM and FPL were the least important covariates for sand at 40-70 cm.

2.5. Discussion

A literature review of DSM efforts reveals that the quantification of the spatial variability of soil texture lacks consensus in terms of the number of samples required for the generation of soil texture maps. There are also a multitude of potential covariates that can be used in this effort and depending on the region and scale of study, different covariates have been found to be useful. The objective of this study was to evaluate the optimal sample size and suitable covariates for a floodplain in the Piedmont region of Georgia. The Georgia Piedmont was chosen because it is an essential area of crop production of cotton and tobacco and more. The poultry industry is also important in the Piedmont. The research location is a research farm located in the Piedmont near a floodplain.

I found 35-45 samples is optimal for the 40-acre field in this study based on the confidence variability of the violin plots and median R^2 and RMSE percentages. This would approximate to about one sample per acre or two samples per hectare. The optimal sample size

was decided by varying the sample sizes used for prediction until the incremental reduction in RMSE from 7% or less and increase in R^2 from 0.80 or higher was not very large . Our RMSE for 35 samples was 4% at 0-10 cm and 6% for 40-70 cm for clay. RMSE for sand was 4.5% at 0-10 cm and 7% at 40-70 cm. These RMSE values are comparable to what has been reported in previous studies. Liao et al. (2013) had RMSE values of 10.65% and 6.90% for sand using kriging and cokriging based on 58 samples. For clay, they reported 5.55% for kriging and 4.74% for cokriging. Zhang and Hartemink (2021) used cubist on their study with 99 samples and reported smaller RMSE values of 2.67%, 5.20%, 4.00%, and 2.48 g kg^{-1} for clay, sand, silt, and total carbon. A comparison of this work with previous studies indicates that while it is possible to get lower RMSE values with more sample sizes, it is not always the case. The performance of DSM models is affected by the choice of algorithm for texture prediction, and it is possible to achieve lower RMSE using fewer samples if different models are selected.

The number of samples that we recommend in this study is less than recommendations made by Wetterlind (2010) who recommended 1.5 samples per ha. Conversely Lai et al. (2021) suggested that a common approach to agricultural sampling is one sample per hectare, which Lai et al. (2021) stated that even for a large field of 50 ha would not provide an adequate sample size to predict texture. Long et al. (2018) suggests 7.5 samples per ha in their precision agriculture study using visible and near-infrared spectroscopy in the 400–2200 nm spectral range to predict soil organic carbon (SOC), plant available [Mg, P, K], pH and texture at farm scale.

While we exhaustively utilized topographic covariates, it is possible that better RMSE values could have been achieved by incorporating more spectral covariates obtained from remote sensing. Unfortunately, given the scale of our study satellite-based remote sensing products utilized in many previous studies were not viable. Liao et al. (2013), used remote sensing data

from Landsat ETM with different bands in a low-density sampling area to get better results for soil texture. For this study, using Landsat bands as covariates was considered but they were found to be too coarse to capture the detailed patterns in the study area. Landsat ETM has a spatial resolution of 30 meters while our study site is 40 acres which would make it challenging to resample and avoid bias, particularly given the 25- fold difference in pixel size for predictions. Hence, the size of the field for which predictions are being made and the desired spatial resolution of covariates can limit the use of covariates.

The spatial distribution of soils in this study site and the dominant covariates can be described jointly based on its location on the landscape and land management. The study area lies in a floodplain. The surface soil gets mixed seasonally and this could be a result of the high clay percentage found in the SW corner and where there are hills. The field has a history of crop production and different plowing practices and mixing soils on the surface can lead to homogeneous soil texture. Topography also affects soil variability by controlling deposition and erosion processes. For this field, finer fractions are usually transported from erosional surfaces and accumulate on the deposit surface such that high silt percentage is found in the flood plain. The field does have variation in topography where the SW corner of the field is generally higher in elevation with clay being exposed near the surface. In the NW corner, the soils tended to be drier than the rest of field and this is where a wash runs through depositing sediments. The NE corner of the field was closer to the Oconee River with the lowest elevation and experienced wetter conditions and intermittent flooding. In the SE section of the filed, there is a hill with more exposed clay closer to the surface.

Irrigation practices such as pivot irrigation are used in the field that can cause water to run off in floodplains and drainage areas since the field is hilly in most areas. There is a wash

that runs from the northwest side of the field to the east side of the field and clay percentage is low in those areas as well. The field is rotated and tilled seasonally. There is a trend where deeper depth increases the percentage of clay, except in floodplains and washes. Clay at 0-10 and 40-70 cm display low percentages at the NE lower floodplain when compared to higher elevation areas (Fig. 3) where clay percentage is high. Higher clay percentage may be due to erosion exposing the B horizon.

Comparing our data to SSURGO, the soil series database describes the field with Chewacla, Wehadkee, and Wickham. The description is not identical but similar to what we predicted. An example is Wickham, which states that the soils are rarely flooded and have high clay content, which is the same as that observed in the prediction maps. On shallow soils (0-10 cm), our predictions are similar to what SSURGO reports. In deeper (40-70 cm) depths, the majority of our predictions differ from SSURGO. The spatial variation of soil properties has different soil texture percentages in the soil series.

Floodplain soils are among the most abundant on Earth due to periodic flooding that deposits nutrient-rich fine-grained sediments (Burwell et al., 1975; “Nutrient Supply to Floodplains | EARTH 111: Water: Science and Society,”). The sediments from the Oconee River are deposited in the lower elevation section of the field that is close to the river. When we auguered soil samples in the areas close to the Oconee River, the surface samples were mainly silt loam. The soil was sandy and saturated at 100 cm depth. As deposition occurs, usually the finer sediment travels and gets deposited as water level decreases. The reason for finding higher amounts of silts as compared to sands and clay in the NE corner can be explained by the location of the study site within the landscape. Soils on active floodplains receive deposits of new alluvium with each flooding episode. The amount of alluvium deposited during each event will

vary. Small amounts of material deposited on the soil can be barely noticeable and quickly incorporated into the underlying surface horizon, whose rate depends on climate and biota. Larger amounts of new alluvium can completely bury underlying soils. This is possibly because of the Wehadkee soil series that lies in that area and is dominated by silt at all depths. Silt has spatial patterns that match the field topography (Figures 19 & 20) especially in the floodplain area that is constantly saturated in the NE section. Silt at 0-10 cm has low percentages in the high elevation areas of the field ranging from 0-32%. Chewacla soils are found on flood plains. They are deep and somewhat poorly drained. They have a brown loamy surface layer and subsoil. The subsoil also contains masses of gray iron depletions which indicate prolonged periods of saturation.

While collecting samples in the Chewacla region of the field, there were lots of iron depletions in the soils. These soils are commonly saturated at depths of 15 cm to 0.61 m (6-24 inches) during the wettest months of the year and are subject to flooding. Chewacla soils formed in sediments washed from the surrounding uplands. Comparing the RF maps, in the Chewacla region, clay is between 10-29% at 0-10 cm and 40-70 cm. Clay content increased with depth. This correlates with the Chewacla soil series description provided by USDA. According to NRCS ("Natural Resources Conservation Service,"), the Wickham soil series have 2 to 6 percent slopes and are rarely flooded. Wickham is found in the Piedmont and Coastal plains in Georgia. The Wickham horizons are generally sandy loam at the surface and three Bt layers. When observing the RF map where Wickham is located, the majority of the soil texture percent is clay. The deeper the depth, the more clay. On the official Wickham soil series description, the 0-15 cm is fine sandy loam. From 15-91 cm, it is all Bt, meaning mostly clay.

Comparing the three-soil series of Chewacla, Weehadkee, and Wickham with environmental covariates and RF prediction maps, we found that there are similar patterns associated with the soil series. Comparing the RF spatial prediction maps with official series description is practically accurate with their descriptions of the soil. For example, for sand at 0-10 cm with 35 samples, the strongest covariate predictor was PCA. Comparing the PCA pattern with the RF map is very similar. But at 40-70 cm for sand, strongest covariate predictor was ECa but the RF prediction map does not show any similar patterns with ECa. Clay at 0-10 cm at 35 samples strongest covariate predictor was FPL. The patterns of FPL have high values in the Wickham and that is where the majority of clay is found on the surface of the field. At 40-70 cm clay, the strongest covariate predictor was DEM at 35 samples. When comparing the official series description for Chewacla that states at 10-152 cm, it is clay dominated which is what the RF maps predicted.

Some limitations with our study were that we sampled by depth and not genetic horizons. This could impact our work because soil texture often varies by horizon and specified depth sampling may have resulted in averaging of soil from contrasting soil textures. An accurate and objective description of the soil profile depends on the identification and exact description of soil horizons (Hartemink and Minasny, 2016). But Hartemink et al., (2020) states that several studies have presented short-range variation and within-horizon variation and have stated that a single core sampling is not sufficient. Several soil cores within a short distance should be collected to capture the short-range variation.

Another limitation was coarse data was available and it was a challenge to find current small-scale resolution for these maps. We had a drone capture imagery of the field for RGB and lidar, butcomplications with the data precluded their use in this study.

Validation and model accuracy are important when making DSM. There are many examples in the literature using machine learning to predict soil attributes with many using leave-one-out cross validation (Wadoux et al., 2019). To make sure the RF model was accurate we followed some fundamental recommendations for modeling. First, we collected new soil samples and analyzed them in the laboratory. This was beneficial for data for the field. Piikki et al. (2021) reported out of 188 DSM literature modelling studies, only 13% of the studies used a type of probability sampling making unbiased estimations of map accuracy. The remaining 87% used soil samples that already have data from multiple samplings. Secondly, we used a robust sample design by applying the cLHS algorithm to obtain a full representation of multivariate distribution in geographical space. Biswas and Zhang (2018) did a review on sample designs in DSM studies and reported out of 95 DSM studies, 15% did not tell any information on their sampling design. Furthermore, we applied some knowledge of the field topography and geomorphology to help decide what topographic covariates to use in the model as variables. And finally, we used cross validation in the model for accuracy using the Caret package in R (Kuhn et al., 2022), which automatically employed tuning parameters for the models.

The choice of sampling design is relevant to the success of a DSM effort. However, it is also subjective to the scale of the study. The cLHS sampling design was used in this study for sample point locations. The cLHS performed a good sampling design of point collection based off the three environmental covariates used within the RF model. This is opposite as to what Wadoux et al. (2019) saw in their case study. They stated that cLHS sampling performs worse than other sampling designs for mapping with RF. Their study site was big with 23 European countries. They also had a sample size of 19,790. They used a freely available soil database within the framework of the European Land Use/Cover area frame Statistical Survey (LUCAS).

The number of environmental covariates in their study was 197 at 1 km x 1 km resolution. Wadoux et al. (2019) does mention that RF benefits from a uniform spread of the sampling locations in feature space of the most important covariates and cLHS was the worst of all other sampling designs. They mention that further research should be considered for cLHS and whether it is good for RF mapping. It is likely that cLHS performs better at creating a sampling design in smaller sizes rather than 23 countries. However, larger areas with more points should be investigated.

Regardless of specific improvements with respect to sampling size and sample design modeling, other places in the world have limitations regarding sample size and data transformation, and incompetence to deal with non-linear relationships as noted in Dash et al. (2022). Appropriate strategies need to be considered for obtaining soil maps with higher prediction accuracies.

This study has succeeded in creating digital soil maps of adequate quality and resolution of recommended soil texture maps suggesting 2 m or less lidar resolution and multispectral bands less than 5 m . Zhang and Hartemink (2021) and Lai et al. (2021) have raised the importance of increasing sample sizes to get better results. While this is true, the purpose of this study was to find the optimal sample size on a 40-acre farm situated on a floodplain in the Piedmont in Georgia. Not much research has been done at this scale in the Georgia Piedmont. Ike and Clutter (1968) examined the variability of forest soils in northeast Georgia and suggested that careful sample designs and sampling quantities are needed to make spatial soil maps.

For future work, comparing different models with different sampling design methods would be beneficial to compare the prediction accuracy. Remote sensing covariates such as surface reflectance bands could be valuable at the small field scale. Sampling by horizon would

be beneficial for pedogenesis, classification and microtopography using DSM to see the change in other soil properties such as soil organic carbon and pH.

2.6. Conclusions

Having accurate soil characterization from a minimum sample size to map spatially variable soil properties is necessary for various agricultural and environmental applications. Determining an optimal sample size is crucial because soil sampling can be expensive and laborious. Thus, the development of a model that is simplified can translate reliable estimation of soil texture with fine resolution. In this study, RF models were developed from validation samples by calculating the correlation between observed and estimated values based off R^2 and RMSE. Fifty iterations were run by selecting various sub-samples from the dataset. Based on the models, majority of sample sizes less than 35 had lower R^2 and RMSE percentages which for clay averaged 0.78 for R^2 and 6% for RMSE. Sand RMSE averaged percent was 6.75% and R^2 was 0.74. Models developed with >35 samples had R^2 averages with ≥ 0.80 and an average RMSE of 5.75% for clay and R^2 of ≥ 0.79 and RMSE at 6% for sand. It was determined that the optimal sample size for this 40-ac field was 35. In addition to producing a customized sample density recommendation for this particular field, results of this study also present a viable method for assessing the performance of similar modeling efforts in other locations in the Georgia Piedmont. This approach for choosing an appropriate sample size for DSM will benefit farmers, stakeholders, and engineers by quantifying the uncertainty that may exist for similar mapping activities. Future work should better quantify the potential uncertainty influenced by sample size in different landscapes to facilitate better recommendations for developing detailed soil property maps to be used in precision agriculture, erosion prediction, watershed modeling, and other intensive land uses.

2.7. References

Amirian-Chakan, A., B. Minasny, R. Taghizadeh-Mehrjardi, R. Akbarifazli, Z. Darvishpasand, et al. 2019. Some practical aspects of predicting texture data in digital soil mapping. Soil Tillage.194.https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db =edselp&AN =S0167198718314302&site=eds-live&custid=uga1

Arrouays, D., L. Poggio, O.A.S. Guerrero, and V.L. Mulder. 2020b. Digital soil mapping and GlobalSoilMap. Main advances and ways forward. Geoderma Reg. 21: e00265. doi: https://doi.org/10.1016/j.geodrs.2020.e00265.

Böhner, J., and T. Selige. 2006. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. SAGA - Anal. Model. Appl. doi: 10.1186/1471-2288-4-5.

Böhner, J., R. Köthe, O. Conrad, J. Gross, A. Ringeler, et al. 2001. Soil regionalisation by means
of terrain analysis and process parameterisation. Eur. Soil Bur.

Boettinger, Janis L., et al., eds. Digital soil mapping: Bridging research, environmental application, and operation. Springer Science & Business Media, 2010b.

Breiman, L. 2001. Random forests. Mach. Learn. doi: 10.1023/A:1010933404324.

Brungard, C.W., and J.L. Boettinger. 2010. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. Digital Soil Mapping.

Brungard, C.W., J.L. Boettinger, M.C. Duniway, S.A. Wills, and T.C. Edwards. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240: 68–83. doi: 10.1016/j.geoderma.2014.09.019.

Carter, Martin R., and E G Greorich. *Soil Sampling and Methods for Analysis*. Lewis, 2008.

Chenghai YANG. High resolution satellite imaging sensors for precision agriculture. Front. Agr. Sci. Eng., 2018, 5(4): 393–405 https://doi.org/10.15302/J-FASE-2018226

Conrad, O., B. Bechtel, M. Bock, H. Dietrich, E. Fischer, et al. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. doi: 10.5194/gmd-8-1991-2015.

Dadfar, H., et al. "Evaluation of a Geonics EM31-3RT probe to delineate hydrologic regimes in a tile-drained field." *Precision agriculture* 12.5 (2011): 623-638.

Duffera, M., J.G. White, and R. Weisz. 2007. Spatial variability of Southeastern U.S. Coastal Plain soil physical properties: Implications for site-specific management. Geoderma 137(3):339.https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edselp&AN=S0016706106002618&site=eds-live&custid=uga1.

Everitt, B. 1992. Book reviews : Chambers JM, Hastie TJ eds 1992: Statistical models in S. California: Wadsworth and Brooks/Cole. ISBN 0 534 16765-9. Stat. Methods Med. Res. 1(2): 220–221. doi: 10.1177/096228029200100208.

Franzen, D. (2018). Soil Variability and Fertility Management. In Precision Agriculture Basics (eds D. Kent Shannon, D.E. Clay and N.R. Kitchen). https://doi.org/10.2134/precisionagbasics.2016.0091.

Ferguson, Richard B, and Gary W Hergert. “Soil Sampling for Precision Agriculture.” Soil Sampling for Precision Agriculture, University of Nebraska Lincoln , 2009, https://extensionpublications.unl.edu/assets/pdf/ec154.pdf.

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

Gaur, Nandita, and Binayak P. Mohanty. "Evolution of physical controls for soil moisture in humid and subhumid watersheds." *Water Resources Research* 49.3 (2013): 1244-1258.

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. 1995. Soil-landscape modelling and spatial prediction of soil attributes. Int. J. Geogr. Inf. Syst. doi: 10.1080/02693799508902047.

Godinho Silva, S.H., P.R. Owens, B.M. Silva, G. César de Oliveira, M. Duarte de Menezes, et al. 2015. Evaluation of Conditioned Latin Hypercube Sampling as a Support for Soil Mapping and Spatial Variability of Soil Properties. Soil Sci. Soc. Am. J. doi: 10.2136/sssaj2014.07.0299.

Guo-Shun, L., J. Hou-Long, L. Shu-Duan, W. Xin-Zhong, S. Hong-Zhi, et al. 2010. Comparison of Kriging Interpolation Precision With Different Soil Sampling Intervals for Precision Agriculture. SOIL Sci. 175(8): 405–415. .

Hengl, T., G.B.M. Heuvelink, and D.G. Rossiter. 2007a. About regression-kriging: From equations to case studies. Spat. Anal. 33(10): 1301–1315. doi: 10.1016/j.cageo.2007.05.001.

Heung, B., H.C. Ho, J. Zhang, A. Knudby, C.E. Bulmer, et al. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265: 62–77. doi: 10.1016/j.geoderma.2015.11.014.

Hummel, J.W., L.D. Gaultney, and K.A. Sudduth. 1996. Soil property sensing for site-specific

crop management. Comput. Electron. Agric. 14(2–3): 121–136. doi: 10.1016/0168-1699(95)00043-7.

Johnston, K., J.M. Ver Hoef, K. Krivoruchko, and N. Lucas. 2001. Using ArcGIS geostatistical analyst. Esri Redlands.

Kerry, R., et al. "Sampling in Precision Agriculture." *Geostatistical Applications for Precision Agriculture*, 2010, pp. 35–63., https://doi.org/10.1007/978-90-481-9133-8_2.

Khaledian, Y., and B.A. Miller. 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. 81: 401–418. doi: https://doi.org/10.1016/j.apm.2019.12.016.

Kidd, D., R. Searle, M. Grundy, A. McBratney, N. Robinson, et al. 2020. Operationalising digital soil mapping – Lessons from Australia. Geoderma Reg. 23: e00335. doi: https://doi.org/10.1016/j.geodrs.2020.e00335.

Lagacherie, P., D. Arrouays, H. Bourennane, C. Gomez, and L. Nkuba-Kasanda. 2020. Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. Geoderma 375: 114503. doi: https://doi.org/10.1016/j.geoderma.2020.114503.

Lawrence, P.G., W. Roper, T.F. Morris, and K. Guillard. 2020. Guiding soil sampling strategies using classical and spatial statistics: A review. Agron. J. 112(1): 493–510. doi: https://doi.org/10.1002/agj2.20048.

Lester, J.G., and A.T. Allen. 1950. DIABASE OF THE GEORGIA PIEDMONT. GSA Bull. 61(11): 1217–1224. doi: 10.1130/0016-7606(1950)61[1217:DOTGP]2.0.CO;2.

Levi, M.R., and C. Rasmussen. 2014. Covariate selection with iterative principal component analysis for predicting physical soil properties. Geoderma 219–220: 46–57. doi: 10.1016/j.geoderma.2013.12.013.

Lopez VMD. Natural Resources Conservation Service (NRCS). Salem Press Encyclopedia of Science. 2020. Accessed March 27, 2022. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=ers&AN=89474787&site=eds-live

Lu, S., B. Liu, Y. Hu, S. Fu, Q. Cao, et al. 2020. Soil erosion topographic factor (LS): Accuracy calculated from different data sources. CATENA 187: 104334. doi: 10.1016/j.catena.2019.104334.

Magnetic Flux, Induction, and Faraday's Law | Boundless Physics. https://courses.lumenlearning.com/boundless-physics/chapter/magnetic-flux-induction-and-faradays-law/ (accessed 25 March 2021).

Malone, B.P., B. Minansy, and C. Brungard. 2019. Some methods to improve the utility of conditioned Latin hypercube sampling. PeerJ. doi: 10.7717/peerj.6451.

Max, K. 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28(5). https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edsdoj&AN=edsdoj.2a4b8da853034e4890ecbfab95b3bcdb&site=eds-live&custid=uga1.

Maxwell, A.E., T.A. Warner, and F. Fang. 2018. Implementation of machine-learning classification in remote sensing: An applied review. Int. J. Remote Sens. 39(9): 2784–2817.

McBratney, A.B., M.L. Mendonça Santos, and B. Minasny. 2003b. On digital soil mapping. Geoderma. doi: 10.1016/S0016-7061(03)00223-4.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of State Calculations by Fast Computing Machines. J. Chem. Phys. 21(6): 1087–1092. doi: 10.1063/1.1699114.

Minasny, Budiman, and Alex B. McBratney. "A conditioned Latin hypercube method for sampling in the presence of ancillary information." *Computers & geosciences* 32.9 (2006): 1378-1388.

Mohamed, S. B., E. J. Evans, and R. S. Shiel. "Mapping techniques and intensity of soil sampling for precision farming." *Proceedings of the Third International Conference on Precision Agriculture*. Madison, WI, USA: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, 1996.

Mokarram, M., G. Roshan, and S. Negahban. 2015. Landform classification using topography position index (case study: salt dome of Korsia-Darab plain, Iran). Model. Earth Syst. Environ. 1(4): 40. doi: 10.1007/s40808-015-0055-9.

Moran, M.S., 2000: "Technology and techniques for remote sensing in agriculture," Assoc. Appl. Biol. and Rem. Sens. Soc. Conf. on Remote Sensing in Agriculture; June 26-28, Cirencester, England; p. 1-10.

Peterson, A. M., Helgason, W. H., & Ireson, A. M. (2019). How spatial patterns of soil moisture dynamics can explain field-scale soil moisture variability: Observations from a sodic landscape. Water Resources Research, 55, 4410– 4426. https://doi.org/10.1029/2018WR023329

Pusch, M., A.L.G. Oliveira, J.V. Fontenelli, and L.R. do Amaral. 2021. SOIL PROPERTIES MAPPING USING PROXIMAL AND REMOTE SENSING AS COVARIATE. Eng. Agríc. 41(6): 634–642. doi: 10.1590/1809-4430-eng.agric.v41n6p634-642/2021

Quinn, P., K. Beven, P. Chevallier, and O. Planchon. 1991. The prediction of hillslope flow

paths for distributed hydrological modelling using digital terrain models. Hydrol. Process. 5(1): 59–79. doi: 10.1002/hyp.3360050106.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

R Studio Team. 2015. R Studio.

Refaeilzadeh, P., L. Tang, and H. Liu. 2009. Cross-Validation. Encyclopedia of Database Systems Robinson, A.C., U. Demšar, A.B. Moore, A. Buckley, B. Jiang, et al. 2017. Geospatial big data and cartography: research challenges and opportunities for making maps that matter. Int. J. Cartogr. doi: 10.1080/23729333.2016.1278151.

Rodriguez-Galiano, V., M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71: 804–818.

Roudier, P. 2011. clhs: a R package for conditioned Latin hypercube sampling.

Roudier, Pierre, D. E. Beaudette, and A. E. Hewitt. "A conditioned Latin hypercube sampling algorithm incorporating operational constraints." *Digital soil assessments and beyond* (2012): 227-231.

RUSLE - an online soil erosion assessment tool. http://www.iwr.msu.edu/rusle/lsfactor.htm (accessed 8 December 2021).

Scull, P., J. Franklin, O.A. Chadwick, and D. McArthur. 2003b. Predictive soil mapping: A review. Prog. Phys. Geogr. doi: 10.1191/0309133303pp366ra.

Sena, Nathalie Cruz, et al. “Soil Sampling Strategy in Areas of Difficult Acess Using the CLHS Method.” *Geoderma Regional*, vol. 24, 2021, https://doi.org/10.1016/j.geodrs.2020.e00354.

Shannon, D., Clay, D.E. and Sudduth, K.A. (2018). An Introduction to Precision Agriculture. In Precision Agriculture Basics (eds D. Kent Shannon, D.E. Clay and N.R. Kitchen).

Sheets, Keith R., and Jan MH Hendrickx. "Noninvasive soil water content measurement using electromagnetic induction." *Water resources research* 31.10 (1995): 2401-2409.

Soil Survey Staff. 2018. Gridded Soil Survey Geographic (gSSURGO) Database for the Conterminous United States. U. S. Dep. Agric. Nat. Resour. Conserv. Serv.

Strobl, C., J. Malley, and G. Tutz. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychol. Methods 14(4): 323–348.

https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edo&AN=46988160&site=eds-live&custid=uga1.

Suleiman, A., and G. Hoogenboom. 2007. Comparison of Priestley-Taylor and FAO56 Penman-Monteith for Daily Reference Evapotranspiration Estimation in Georgia. J. Irrig. Drain. Eng.-Asce - J IRRIG DRAIN ENG-ASCE 133. doi: 10.1061/(ASCE)0733-9437(2007)133:2(175).

Webster, R., and M.A. Oliver. 1992. Sample adequately to estimate variograms of soil properties. J. Soil Sci. 43(1): 177–192. doi: https://doi.org/10.1111/j.1365-2389.1992.tb00128.x.

Weights, I.D.W.V.K. Visualization tools to aid in the understanding of geostatistics.

Weiss, a. 2001. Topographic position and landforms analysis. Poster Present. ESRI User Conf. San Diego CA.

Wetterlind, Johanna, Bo Stenberg, and Mats Söderström. "Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models." *Geoderma* 156.3-4 (2010): 152-160.

Wilson, John P., and John C. Gallant. "Digital terrain analysis." *Terrain analysis: Principles and applications 6.12 (2000): 1-27.*

Wollenhaupt, N. 1996. Sampling and testing for variable rate fertilization. Proceedings of the 1996 Information Agriculture Conference. P & PI Norcross, GA. p. 33–34

Wulfsohn, D. "Sampling techniques for plants and soil." Advanced Engineering Systems for Specialty Crops: A Review of Precision Agriculture for Water, Chemical, and Nutrient Application, and Yield Monitoring (2010): 3-30.

Zhu, Qing, Henry Lin, and James Doolittle. "Repeated electromagnetic induction surveys for improved soil mapping in an agricultural landscape." *Soil Science Society of America Journal* 74.5 (2010): 1763-1774.

Zimmerman, D., C. Pavlik, A. Ruggles, and M.P. Armstrong. 1999. An experimental Comparison of ordinary and universal kriging and inverse distance weighting. Math. Geol. 31(4): 375–390.

2.8 Tables

Table 1. Summary statistics of measured soil properties by depth.

0-10 cm	*Clay*	*Silt*	*Sand*
	%	%	%
Min	11.8	26.2	11.7
Max	53.2	58.4	54.9
Median	27.9	40.4	30.4
Mean	28.3	41.0	30.7
Std	8.1	6.7	9.1
Skewness	0.8	0.3	0.4
Range	41.4	32.2	43.2
Kurtosis	4.1	2.7	3.1

10-40 cm	*Clay*	*Silt*	*Sand*
	%	%	%
Min	16.2	21.1	6.6
Max	62	58.7	53.3
Median	40.1	35.7	23
Mean	40.2	36.1	23.6
Std	11.8	8.5	9.5
Skewness	0.02	0.5	0.7
Range	45.8	37.6	46.6
Kurtosis	2.2	3.0	3.9

40-70 cm	*Clay*	*Silt*	*Sand*
	%	%	%
Min	18.1	17.3	6.6
Max	70.9	54.8	63.4
Median	48.9	30.9	19.2
Mean	45.6	31.2	23.0
Std	14.5	8.5	12.3
Skewness	-0.2	0.6	1.1
Range	52.8	37.5	56.7
Kurtosis	2.0	3.1	3.7

70-100 cm	*Clay*	*Silt*	*Sand*
	%	%	%
Min	16.4	15.7	7.6
Max	69.8	56.6	65.5
Median	48.5	29.8	19.6
Mean	46.0	30.4	23.6
Std	15.3	9.5	13.3
Skewness	-0.3	0.6	1.2
Range	53.4	40.9	57.9
Kurtosis	1.9	3.1	3.7

Table 2. Covariates used for the model. Abbreviations for each covariate are given to understand figures of the covariates. References are also made to the description. Covariates with a * represent covariates used in cLHS SD.

	Covariate	***Description***	***Cited***	***Abbreviation***
Terrain	LS Factor	Factor that reflects slope length and steepness effects on erosion	Boehner, J., Selige, T. (2006)	LSF
	Flow Path Length *	The maximum distance of water flow to a location in a catchment	Freeman, G.T. (1991)	FPL
	Aspect	Aspect identifies the downslope direction of the maximum rate of change in value from each cell to its neighbors.	Brewer, C.A. & Marlow, K.A. (1993)	Aspect
	Topographic Position Index *	An algorithm increasingly used to measure topographic slope positions and to automate landform classifications	Weiss, A.D. (2000)	TPI Landform
	Saga Wetness Index*	The 'SAGA Wetness Index' is, as the name says, like the 'Topographic Wetness Index' (TWI), but it is based on a modified catchment area calculation ('Modified Catchment Area'), which does not think of the flow as very thin film. As a result it predicts for cells situated in valley floors with a small vertical distance to a channel a more realistic, higher potential soil moisture compared to the standard TWI calculation.	Boehner, J., Koethe, R. Conrad, O., Gross, J., Ringeler, A., Selige, T. (2002)	SWI
	DEM	Digital elevation model. Representation of elevation data to represent terrain	USGS	DEM
	Multiresolution Index of Ridgetop Flatness	Topographic index designed to identify high flat areas at a range of scales. It complements the MRVBF index that is designed to identify areas of deposited material in flat valley bottoms.	Gallant, J.C., Dowling, T.I. (2003)	MRVBF
	Multiresolution Index of Valley Bottom Flatness	This index classifies degrees of valley bottom flatness, which may be related to depth of deposit. The index can also be used to identify groundwater	Gallant, J.C., Dowling, T.I. (2003)	MRRTF
Remote Sensing	Red and Green Spectral	Red and green spectral imagery done by a drone. The PCA is used as a covariate.		PCA
EM-31	Electromagnetic Conductivity	Electrical conductivity measurements performed on the crop field and used as a covariate		EM

2.9. Figures

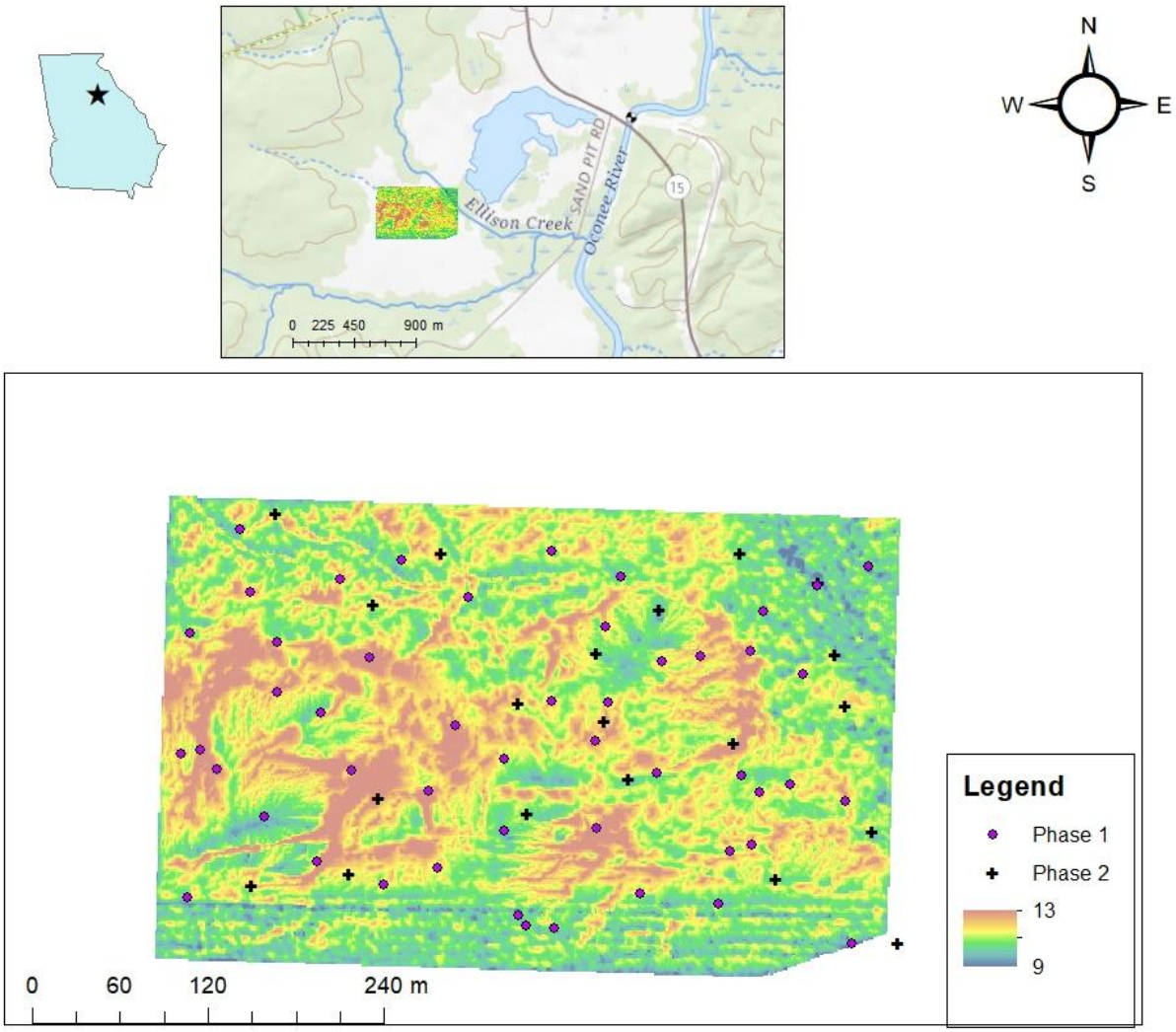

Fig 1. Study location showing soil sample locations. Saga wetness index is shown for variability of the research area. The field has an area in the northeast corner where the soil is saturated most of the time in the floodplain near the Oconee River with less clay and more silt.

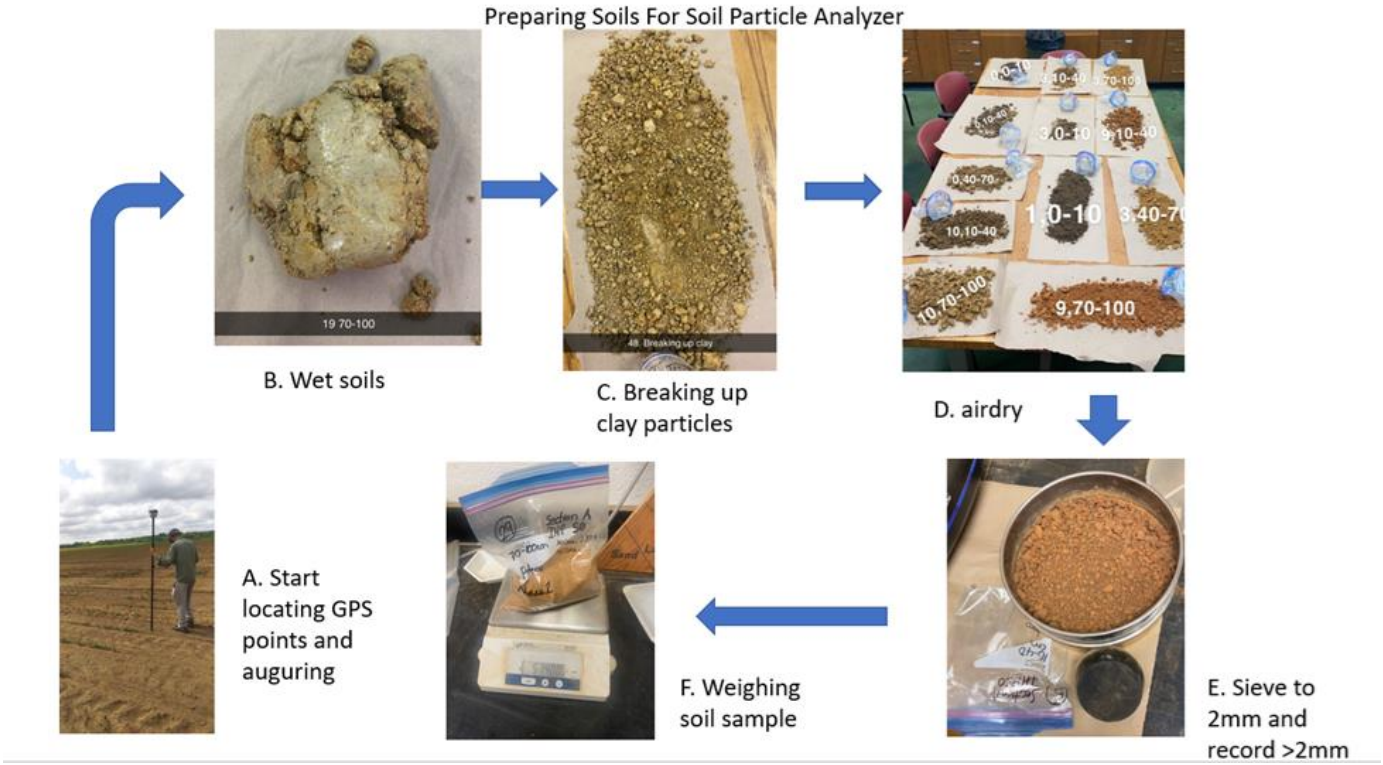

Fig 2. Conceptual diagram of soil sampling approach.

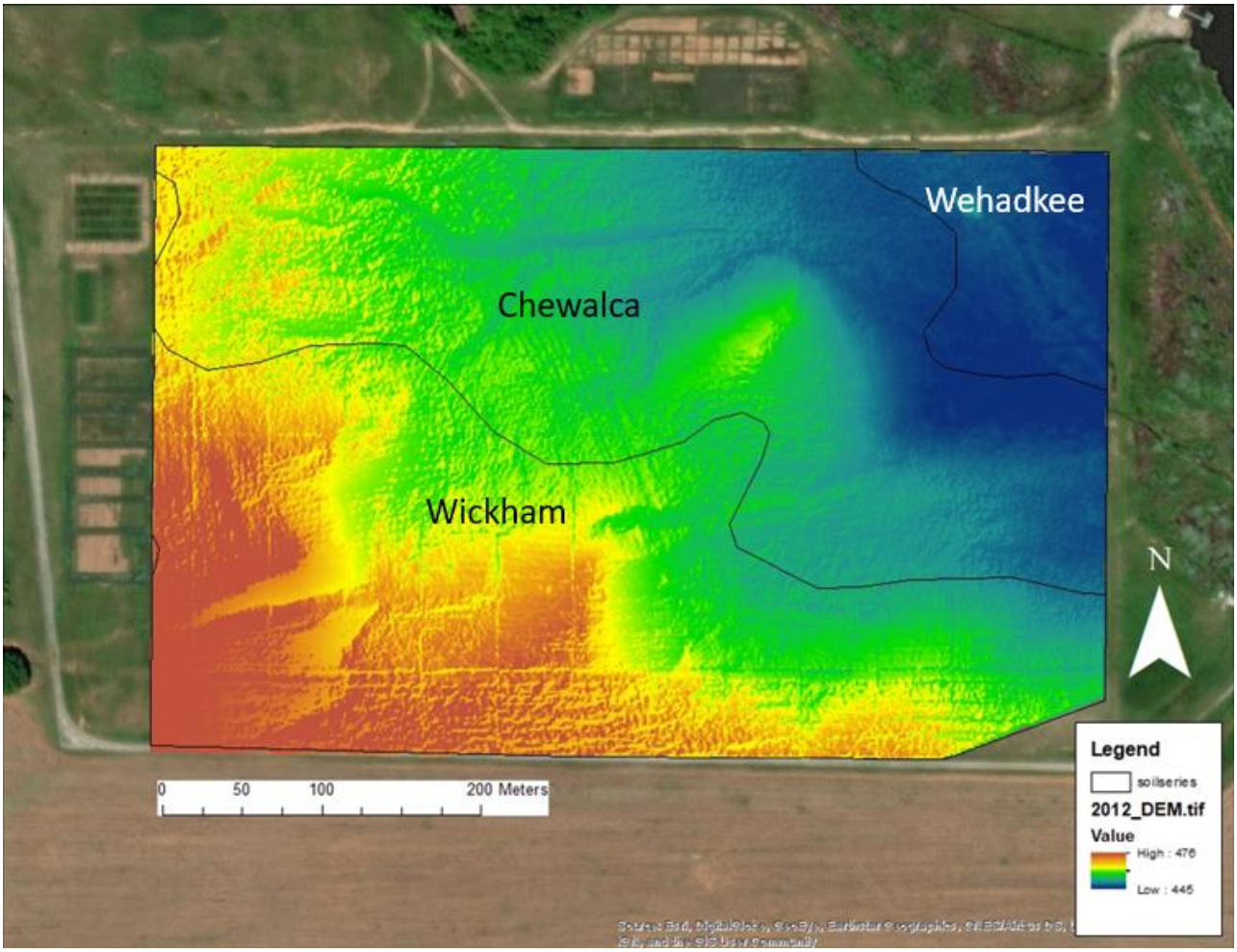

Fig. 3. Map of the field where lower floodplain and upper floodplain are located . Lower floodplain lies in majority of the Wehadkee with low elevation . Elevation increases as colors on the map change towards a reddish orange in the Wickham area.

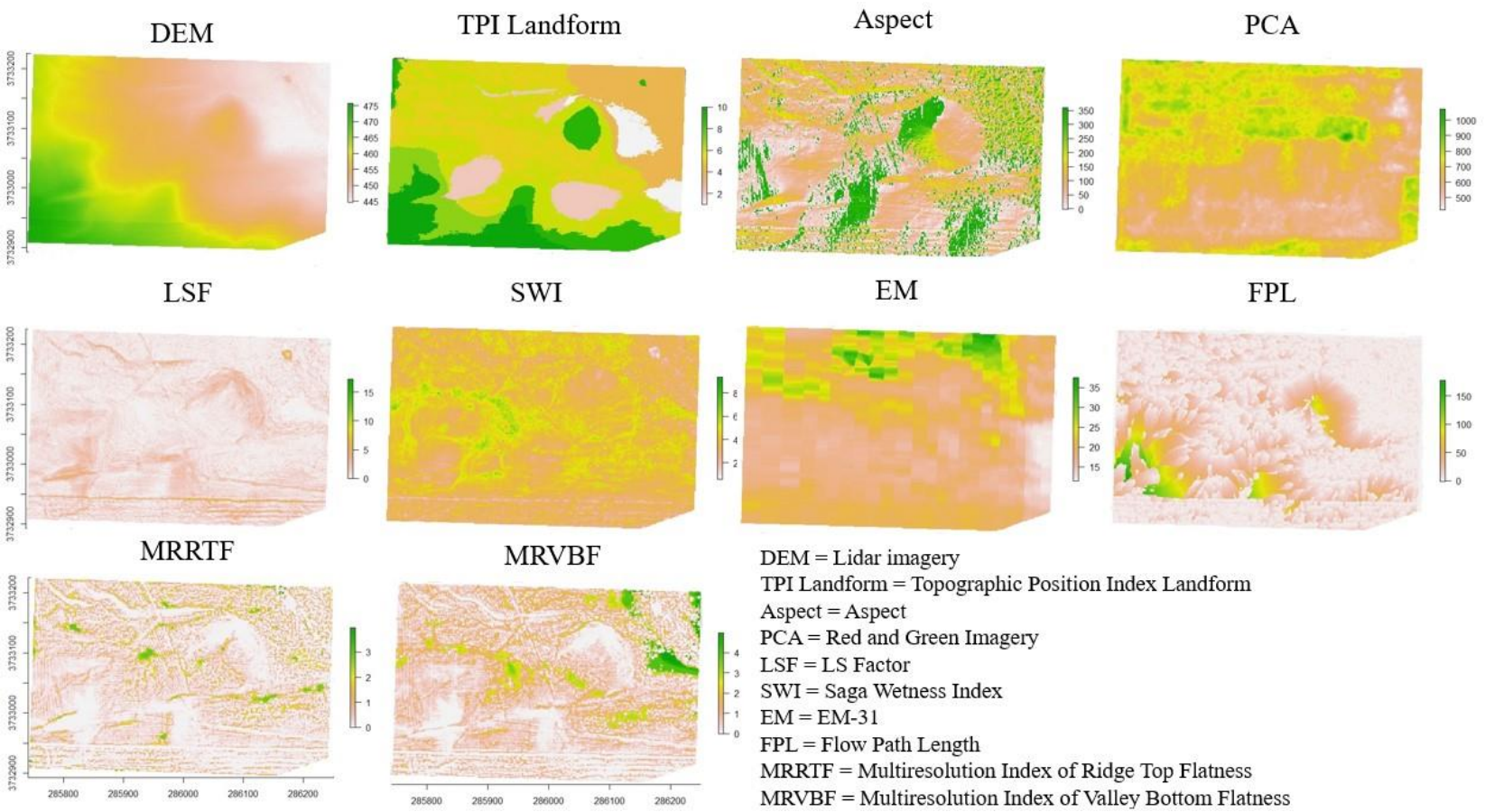

Fig 4. Map of covariates used in random forest models developed at the Iron Horse farm.

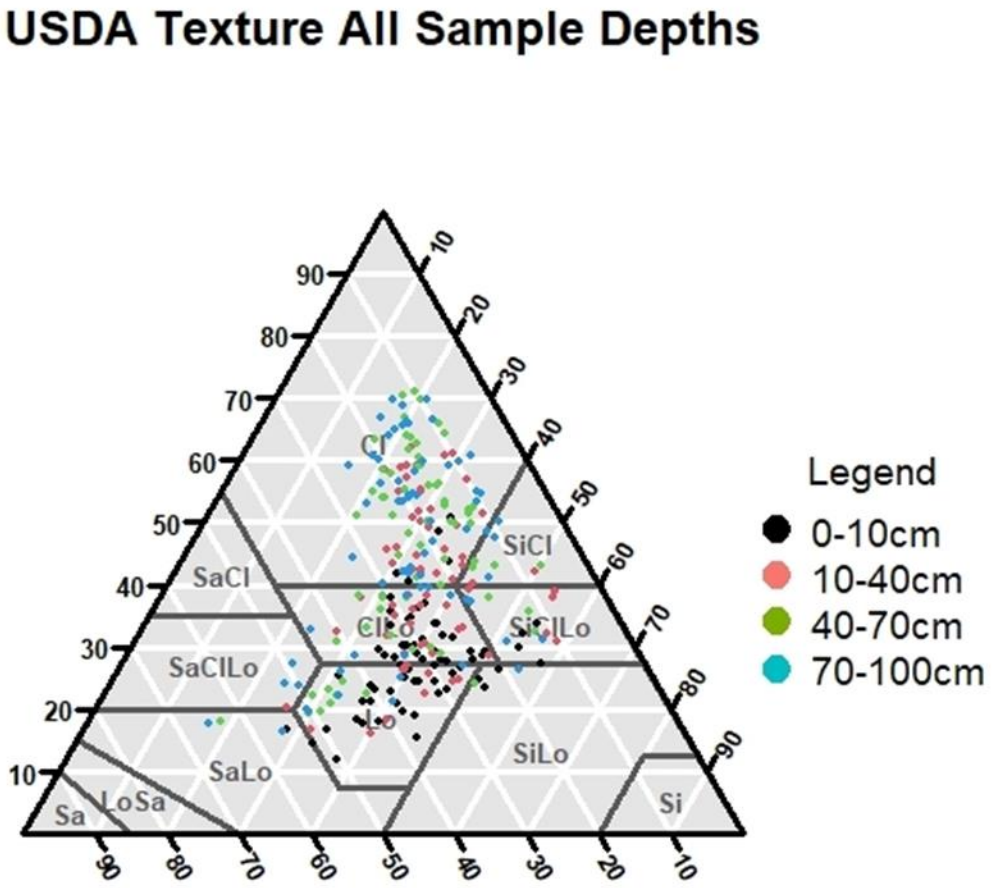

Fig. 5. Soil texture distribution for all measured soil properties. Shallow soils are mainly clay loam and loam. Deeper soils tended to have an increase in clay.

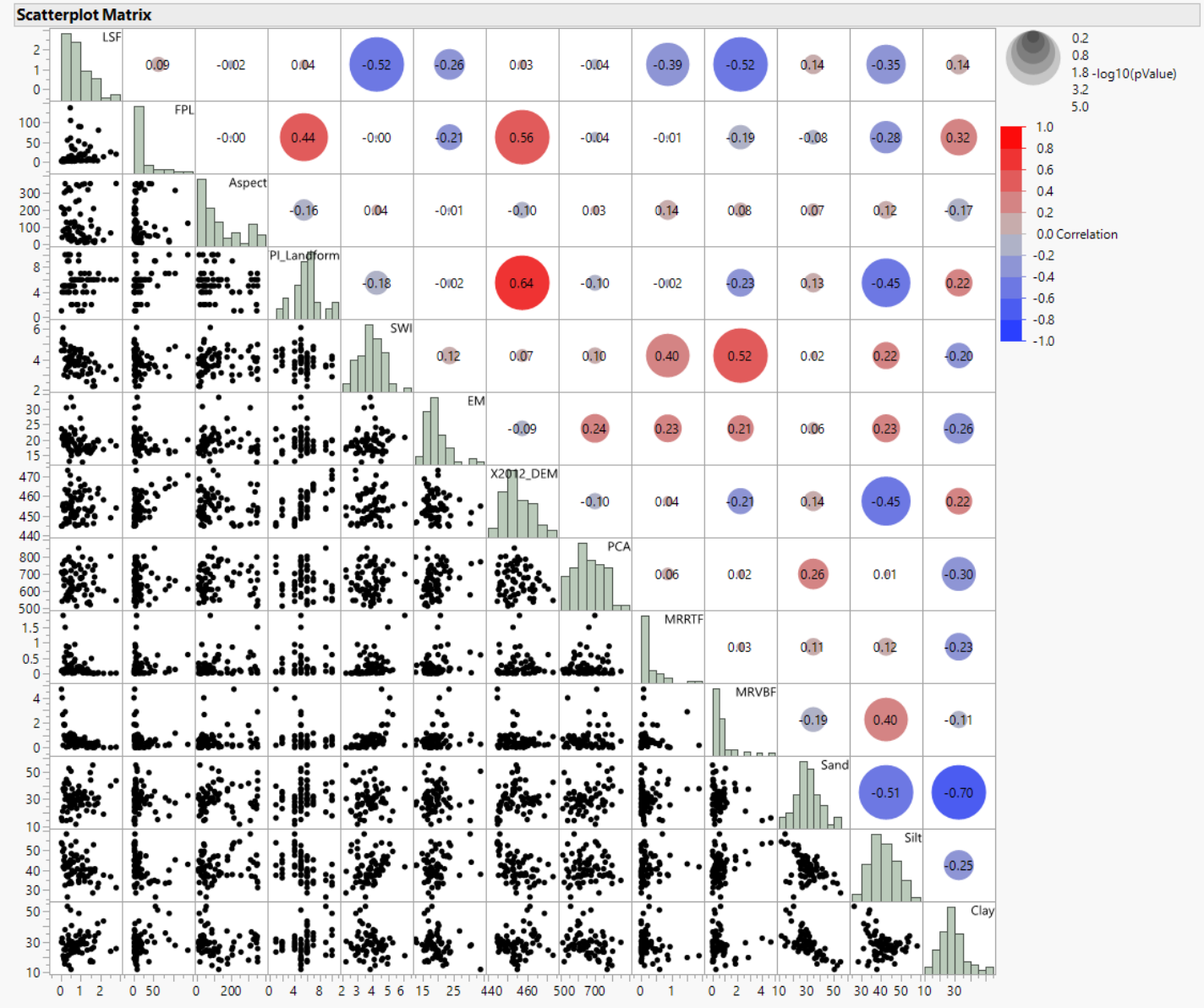

Fig 6. Pearson correlation coefficients among covariates and measured soil texture at 0-10 cm.

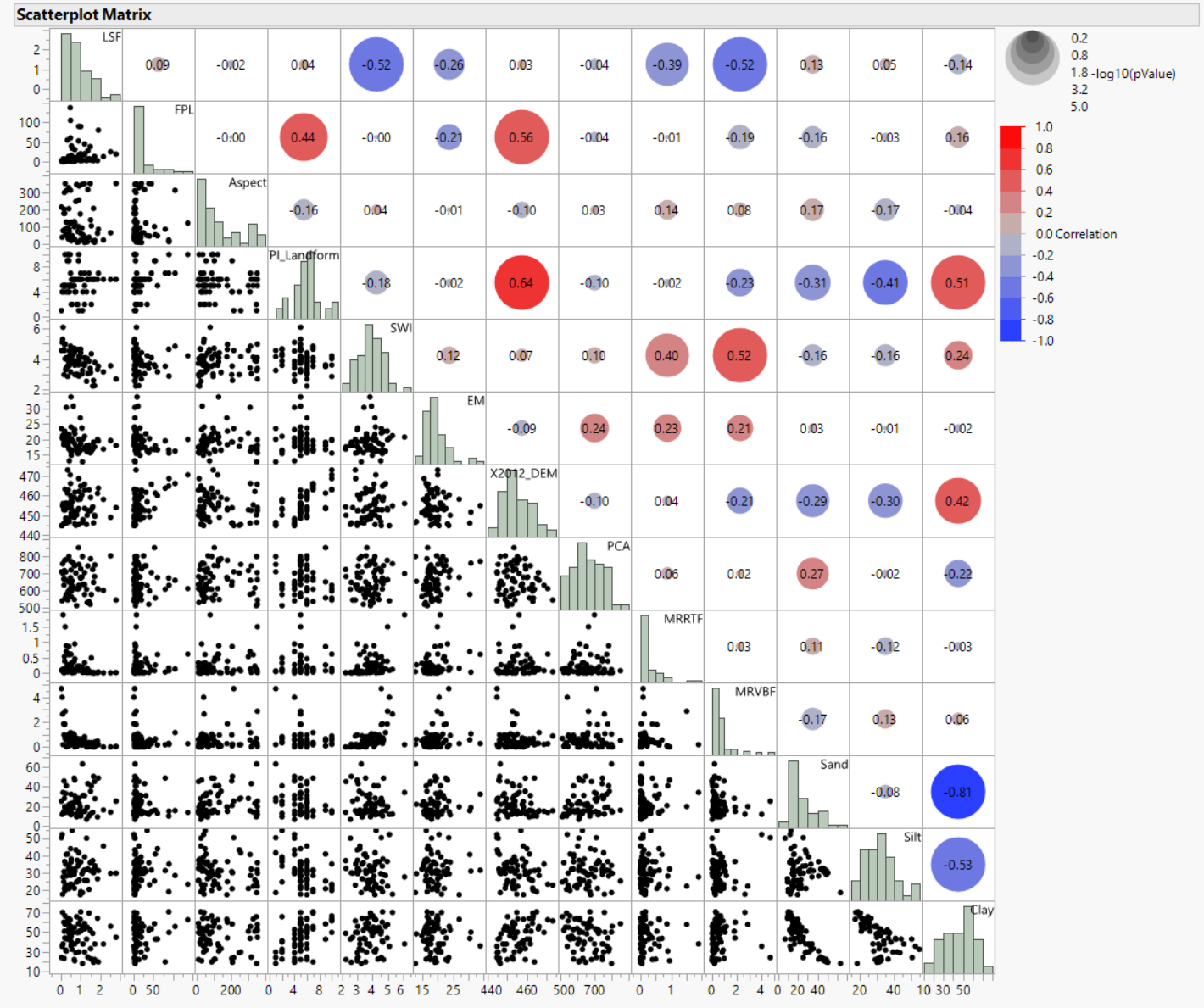

Fig. 7. Pearson correlation coefficients among covariates and measured soil texture at 40-70 cm.

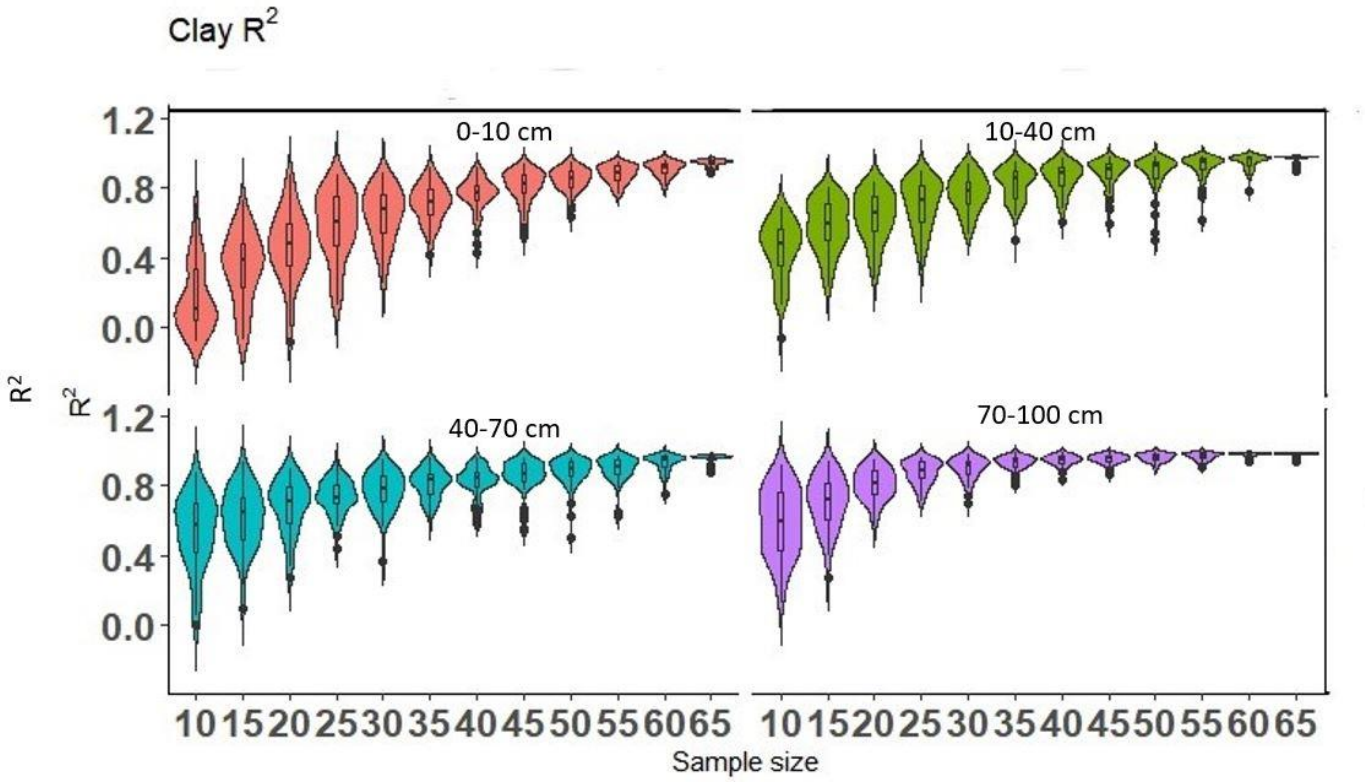

Fig. 8. Model performance (R^2) over 50 iterations for clay at four depths.

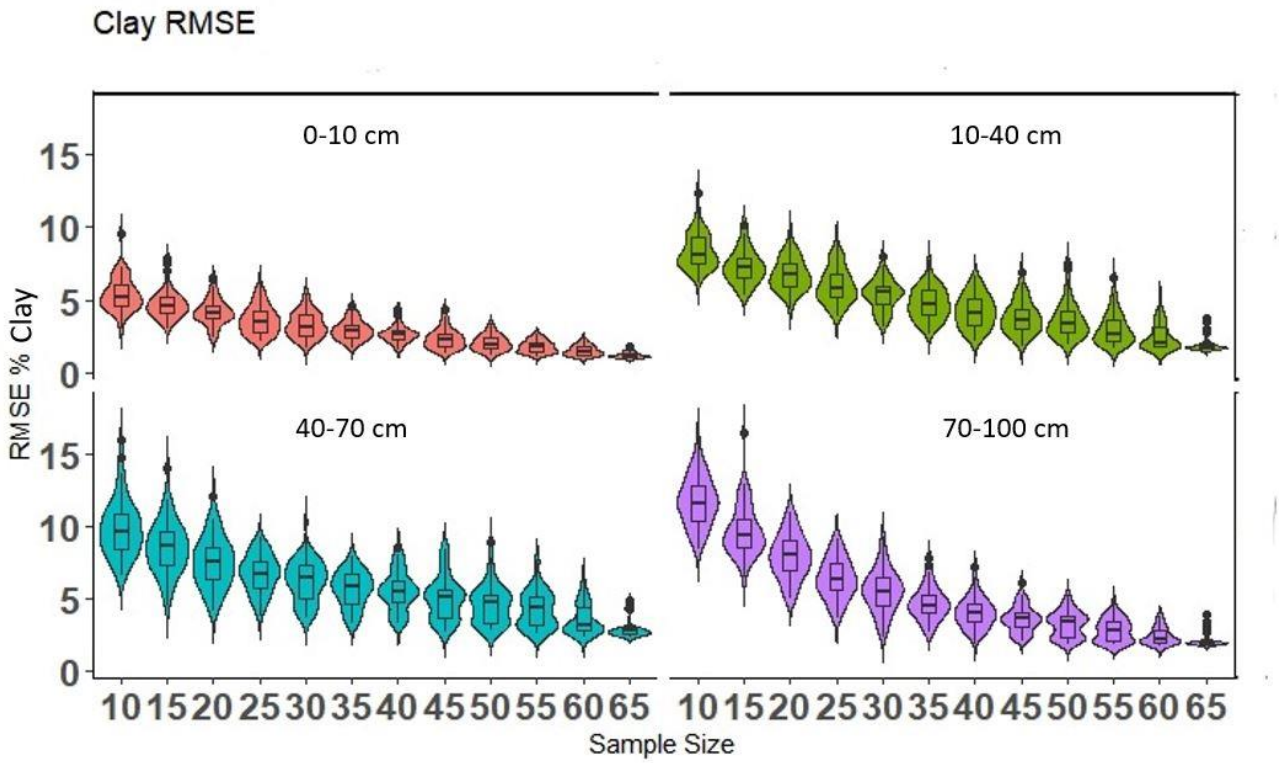

Fig. 9. Model performance (RMSE) over 50 iterations for clay at four depths.

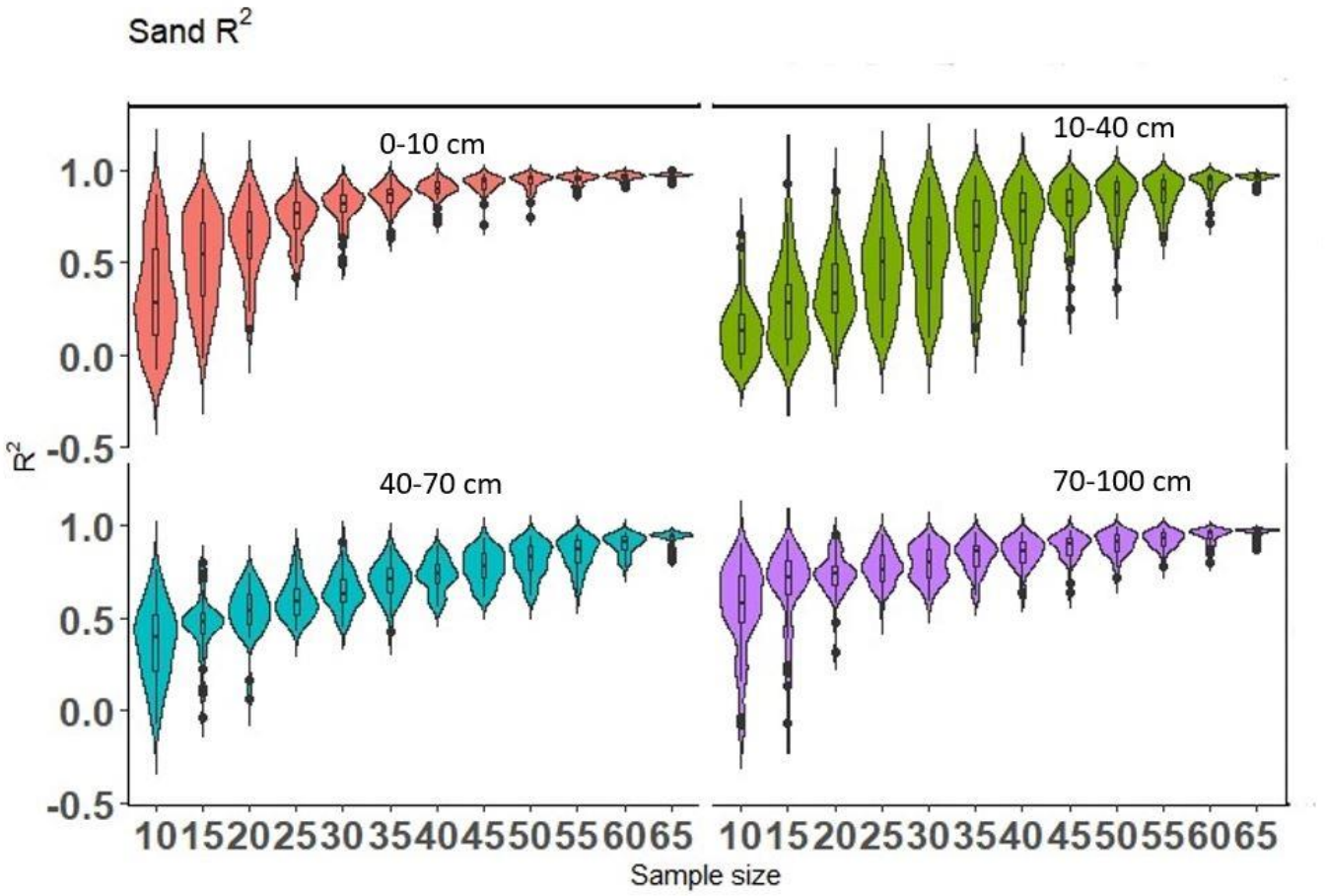

Fig. 10. Model performance (R^2) over 50 iterations for sand at four depths.

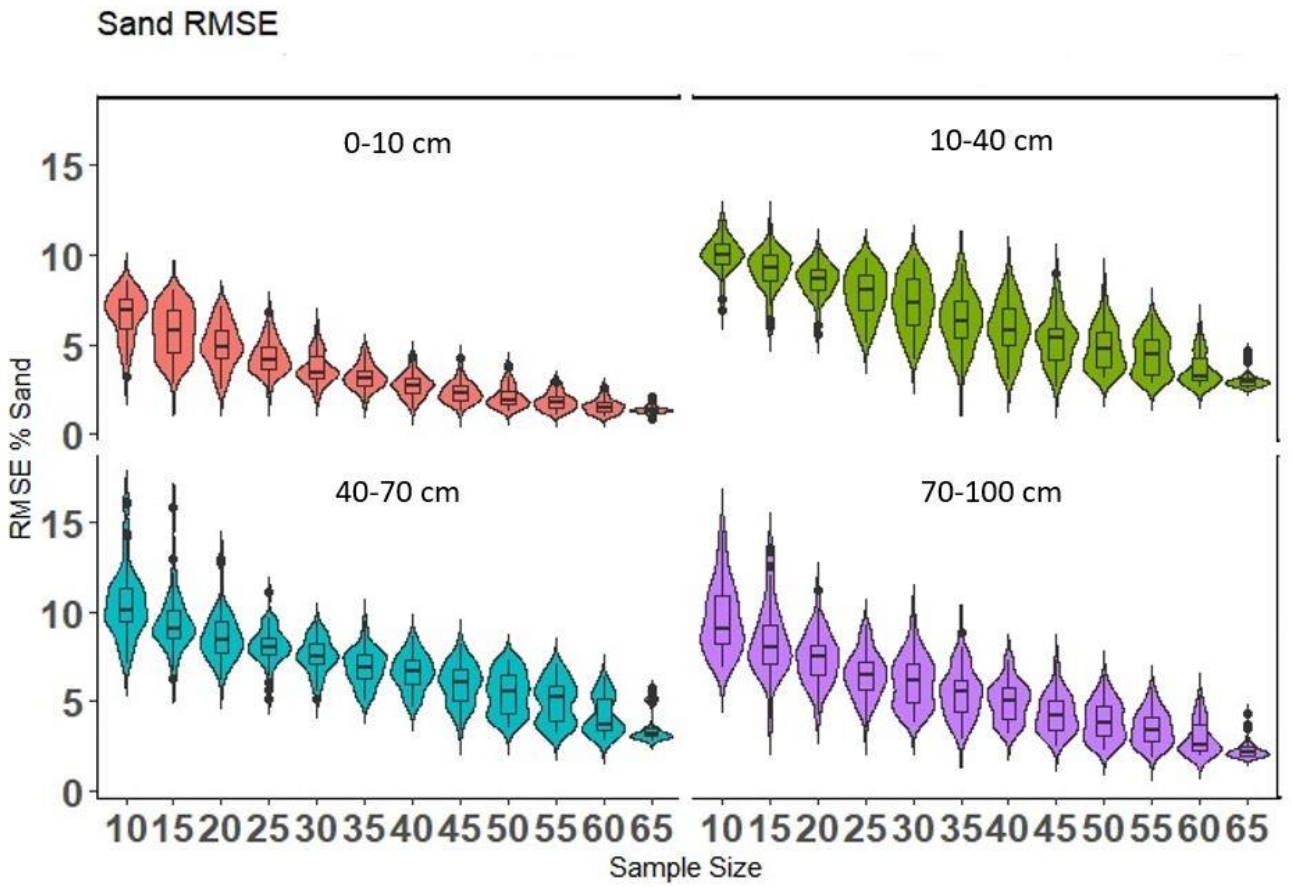

Fig. 11. Model performance (RMSE) over 50 iterations for sand at four depths.

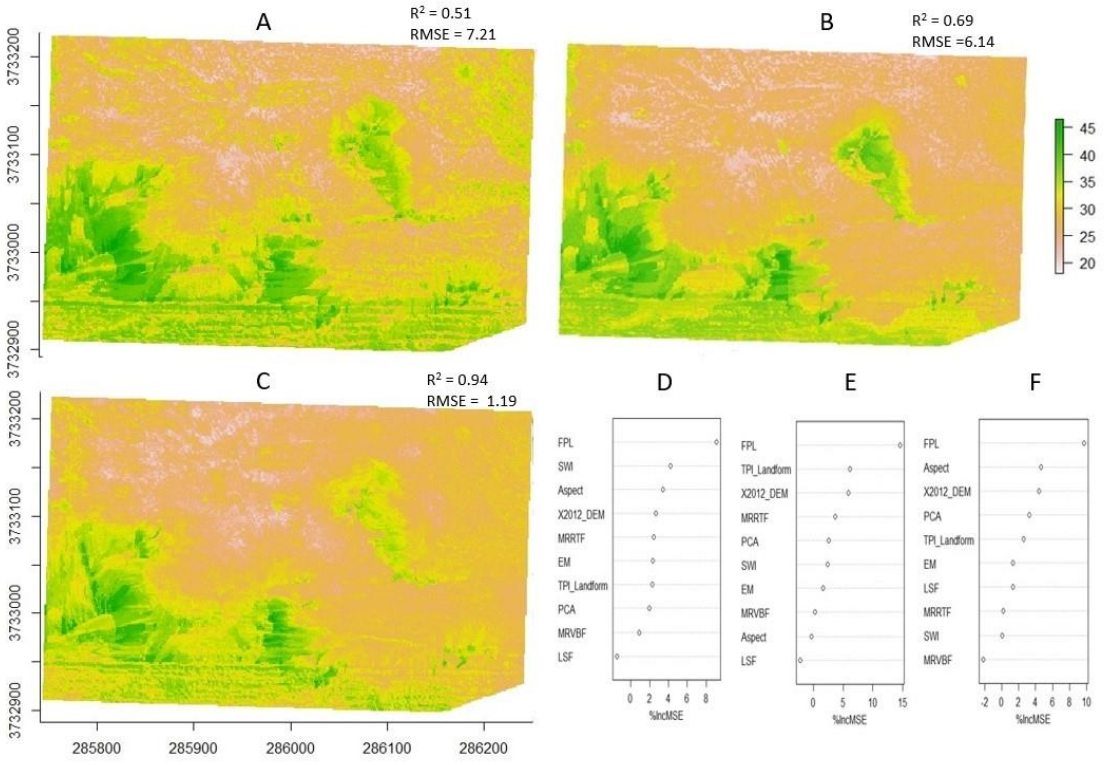

Fig. 12. Predicted clay percentage for 0-10 cm (A, B, C) and variable importance (D, E, F) at 20, 35, and 69 samples, respectively, for one model iteration at each sample size. A, D= 20 samples, B, E= 35 samples, C, F= 69 samples.

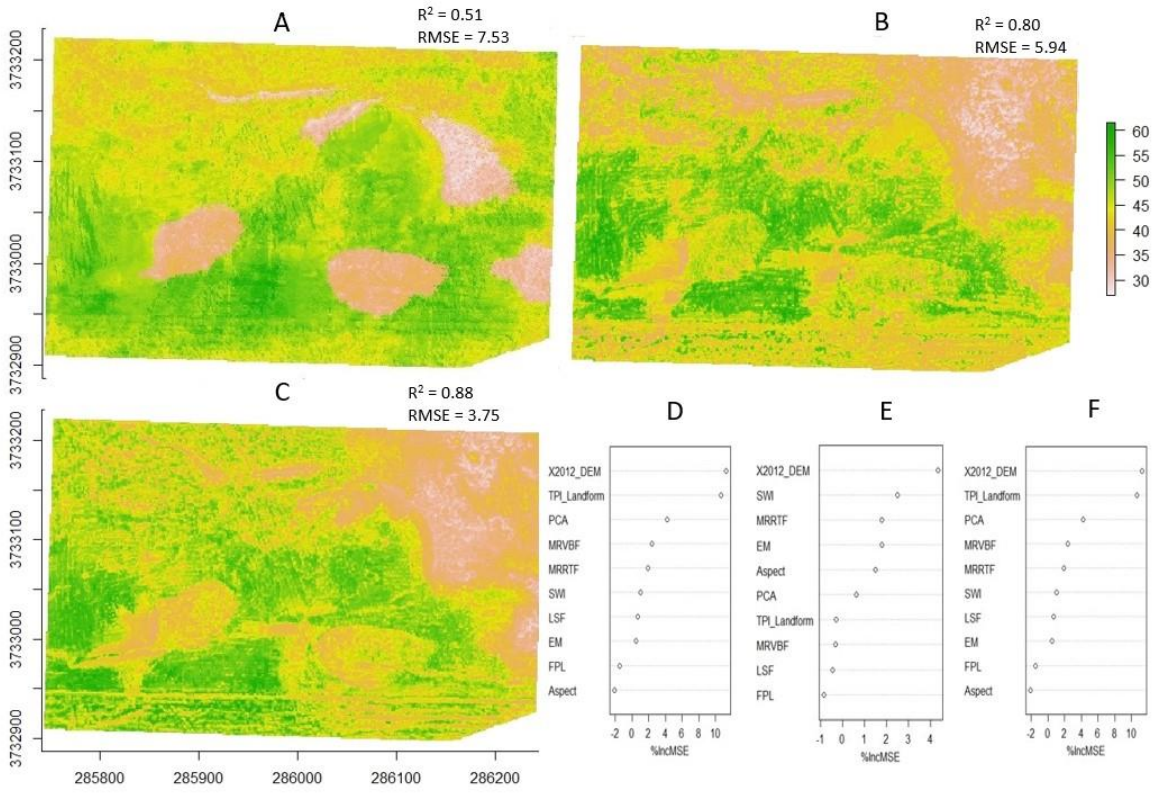

Fig. 13. Predicted clay percentage for 40-70 cm (A, B, C) and variable importance (D, E, F) at 20, 35, and 69 samples, respectively, for one model iteration at each sample size. A, D= 20 samples, B, E= 35 samples, C, F= 69 samples.

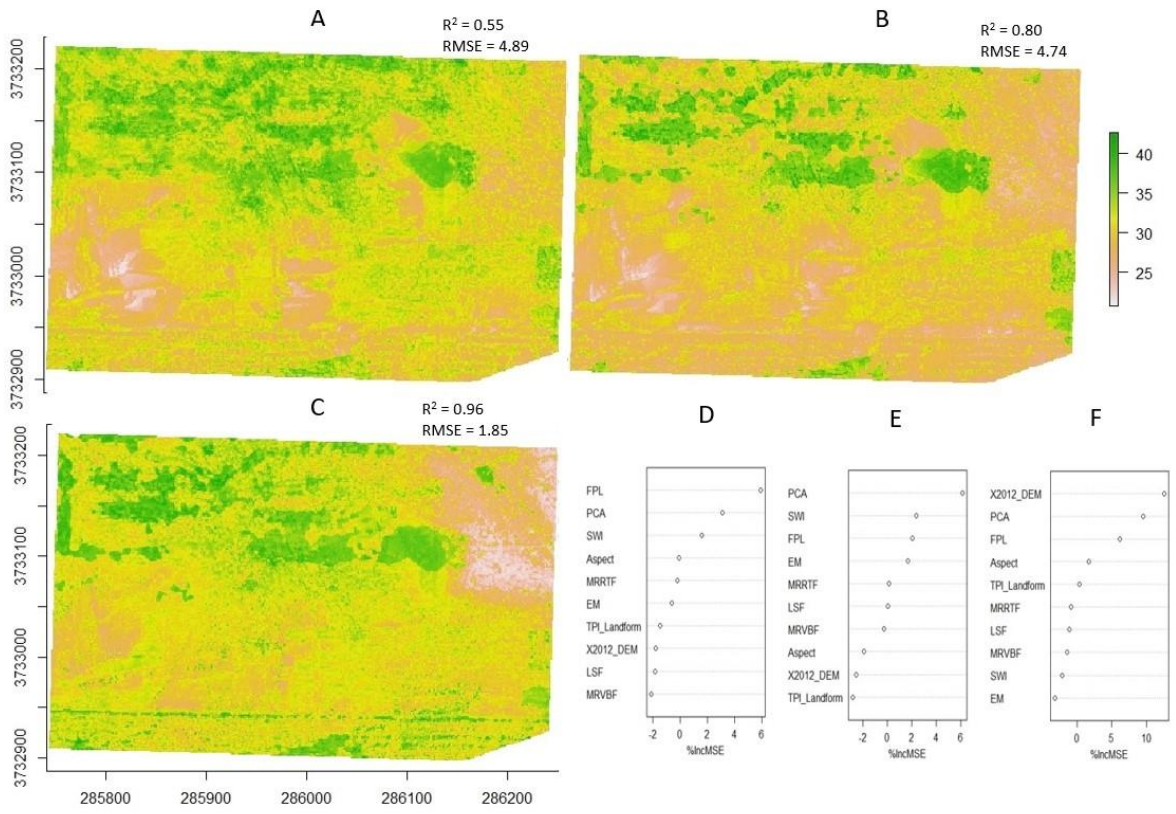

Fig. 14. Predicted sand percentage for 0-10 cm (A, B, C) and variable importance (D, E, F) at 20, 35, and 69 samples, respectively, for one model iteration at each sample size. A, D= 20 samples, B, E= 35 samples, C, F= 69 samples.

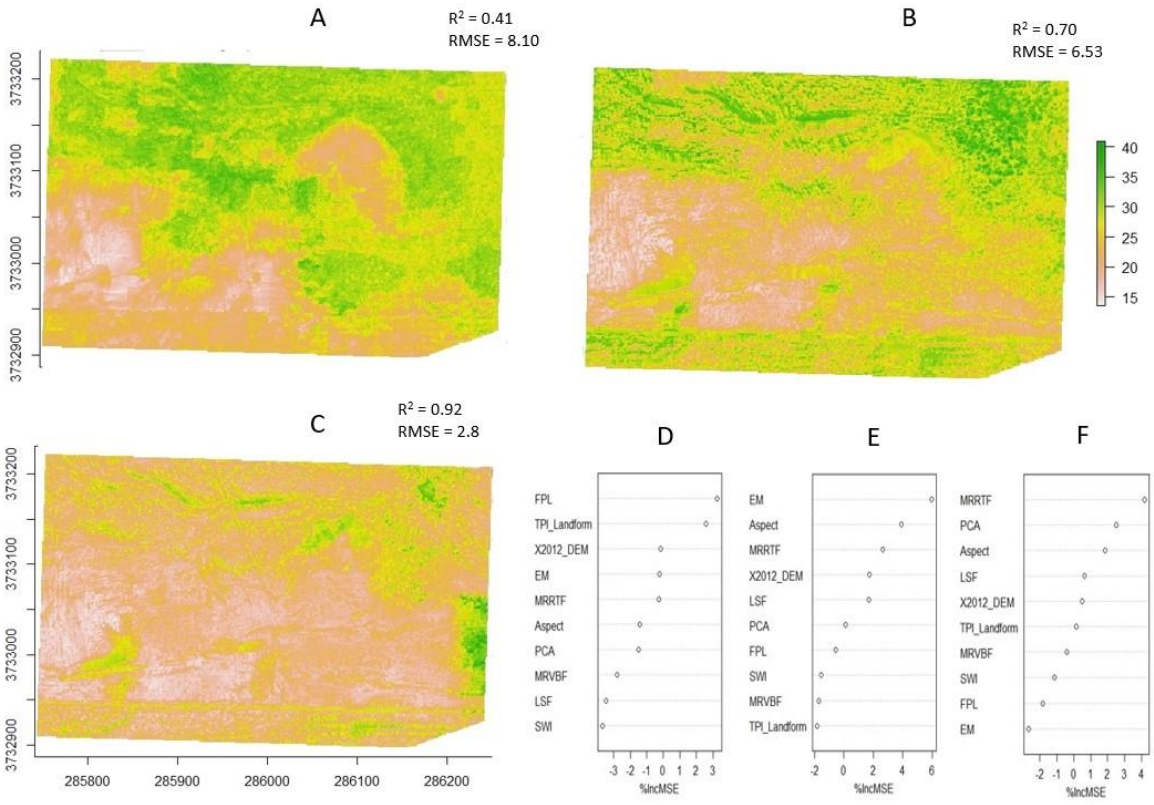

Fig. 15. Predicted sand percentage for 40-70 cm (A, B, C) and variable importance (D, E, F) at 20, 35, and 69 samples, respectively, for one model iteration at each sample size. A, D= 20 samples, B, E= 35 samples, C, F= 69 samples.

Appendix

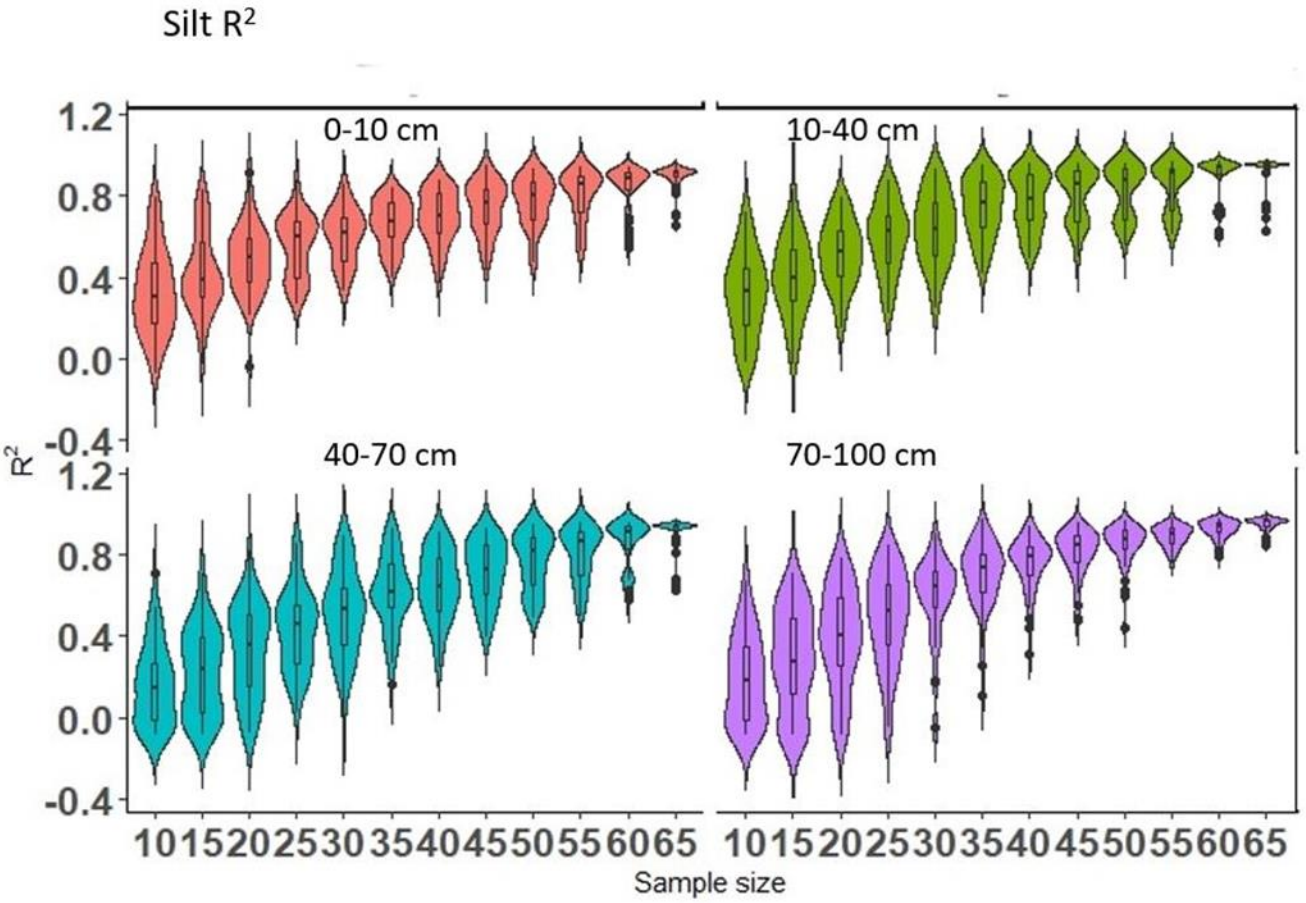

Fig. 16. Model performance (R^2) over 50 iterations for silt at four depths.

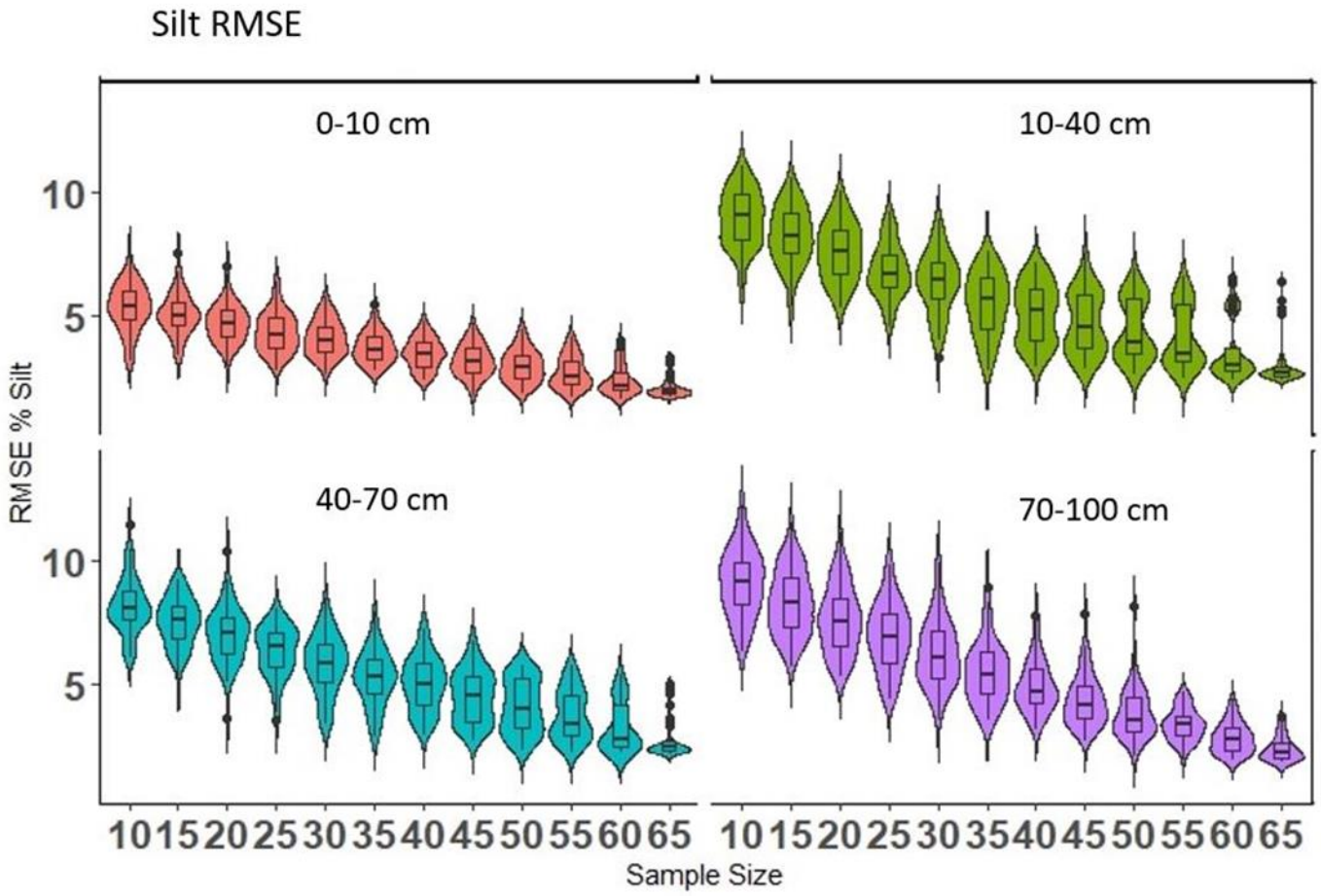

Fig. 17. Model performance (RMSE) over 50 iterations for silt at four depths.

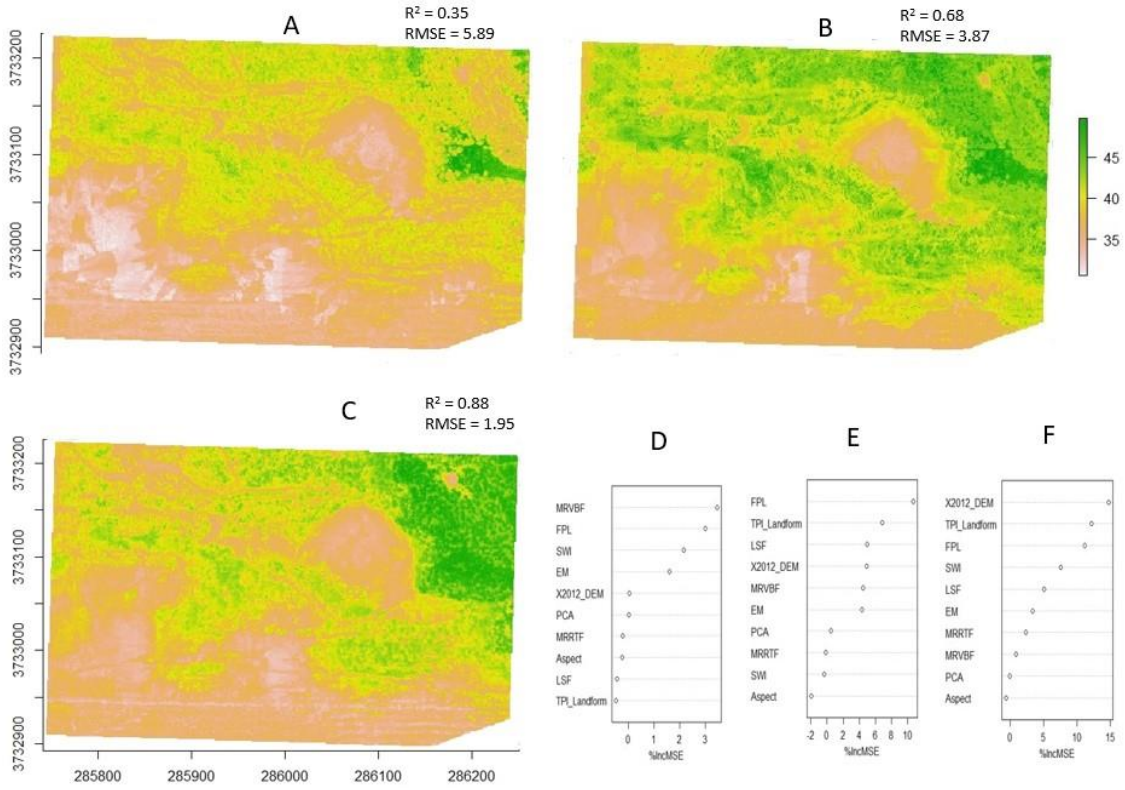

Fig 18. Predicted silt percentage for 0-10 cm (A, B, C) and variable importance (D, E, F) at 20, 35, and 69 samples, respectively, for one model iteration at each sample size. A, D= 20 samples, B, E= 35 samples, C, F= 69 samples.

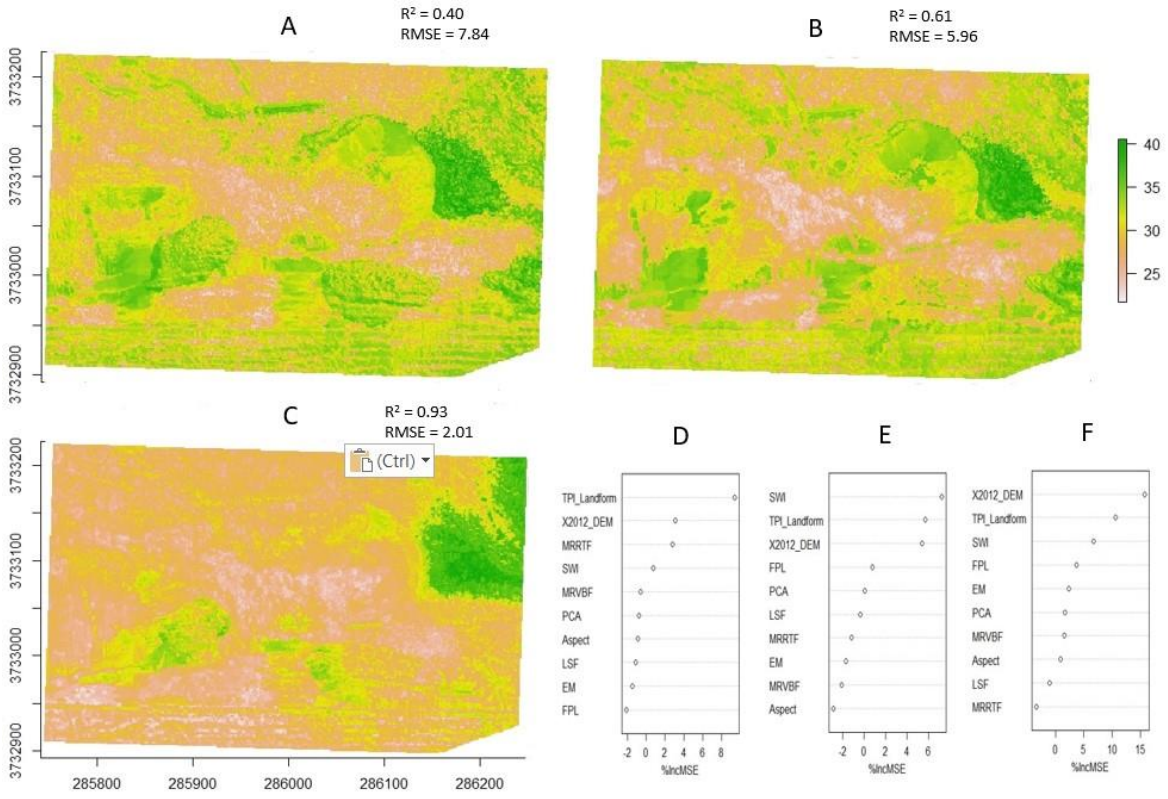

Fig. 19. Predicted silt percentage for 40-70 cm (A, B, C) and variable importance (D, E, F) at 20, 35, and 69 samples, respectively, for one model iteration at each sample size. A, D= 20 samples, B, E= 35 samples, C, F= 69 samples.

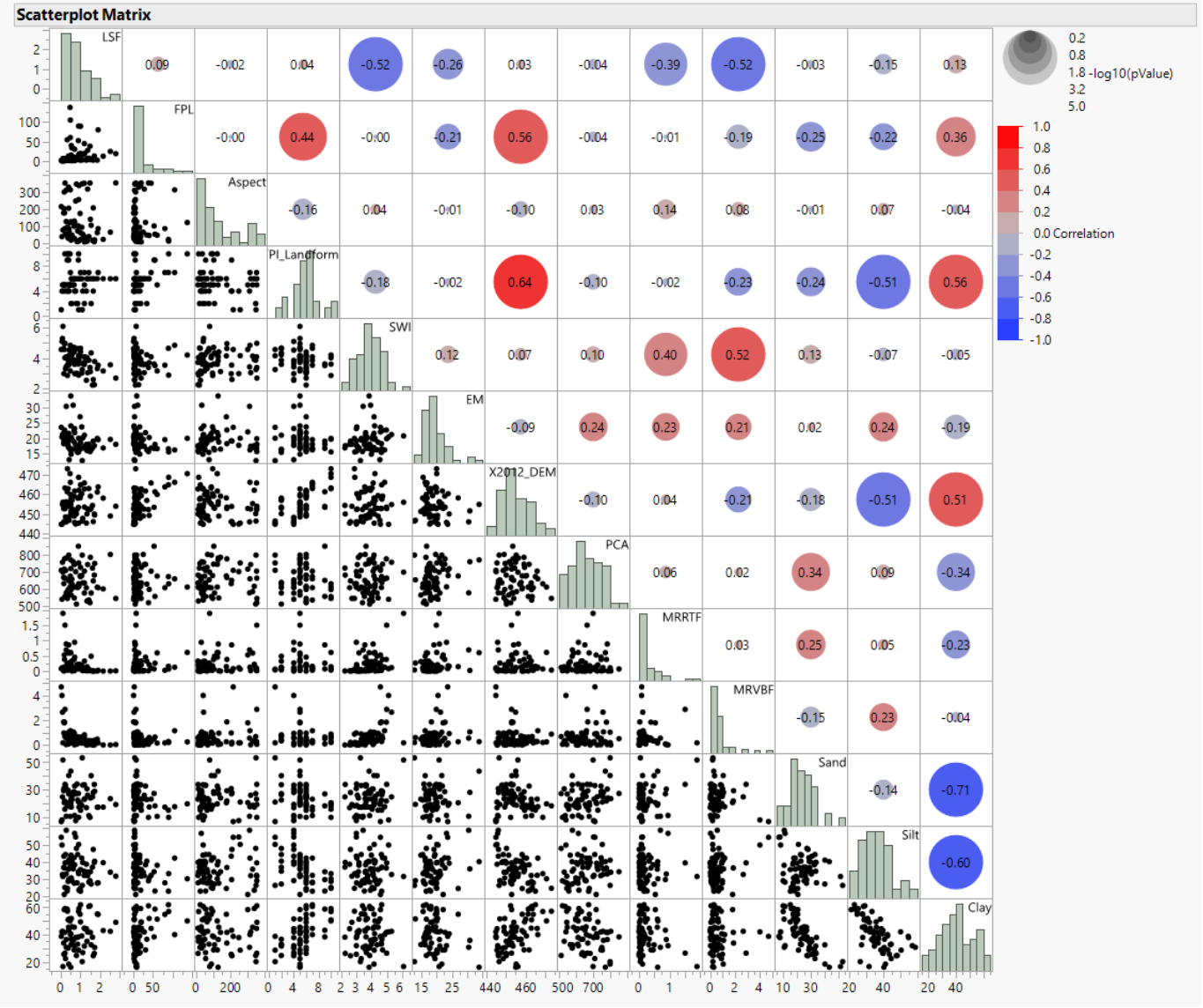

Fig. 20. Correlation matrix and p-values of measured sand, silt, clay at 10-40 cm and environmental covariates.

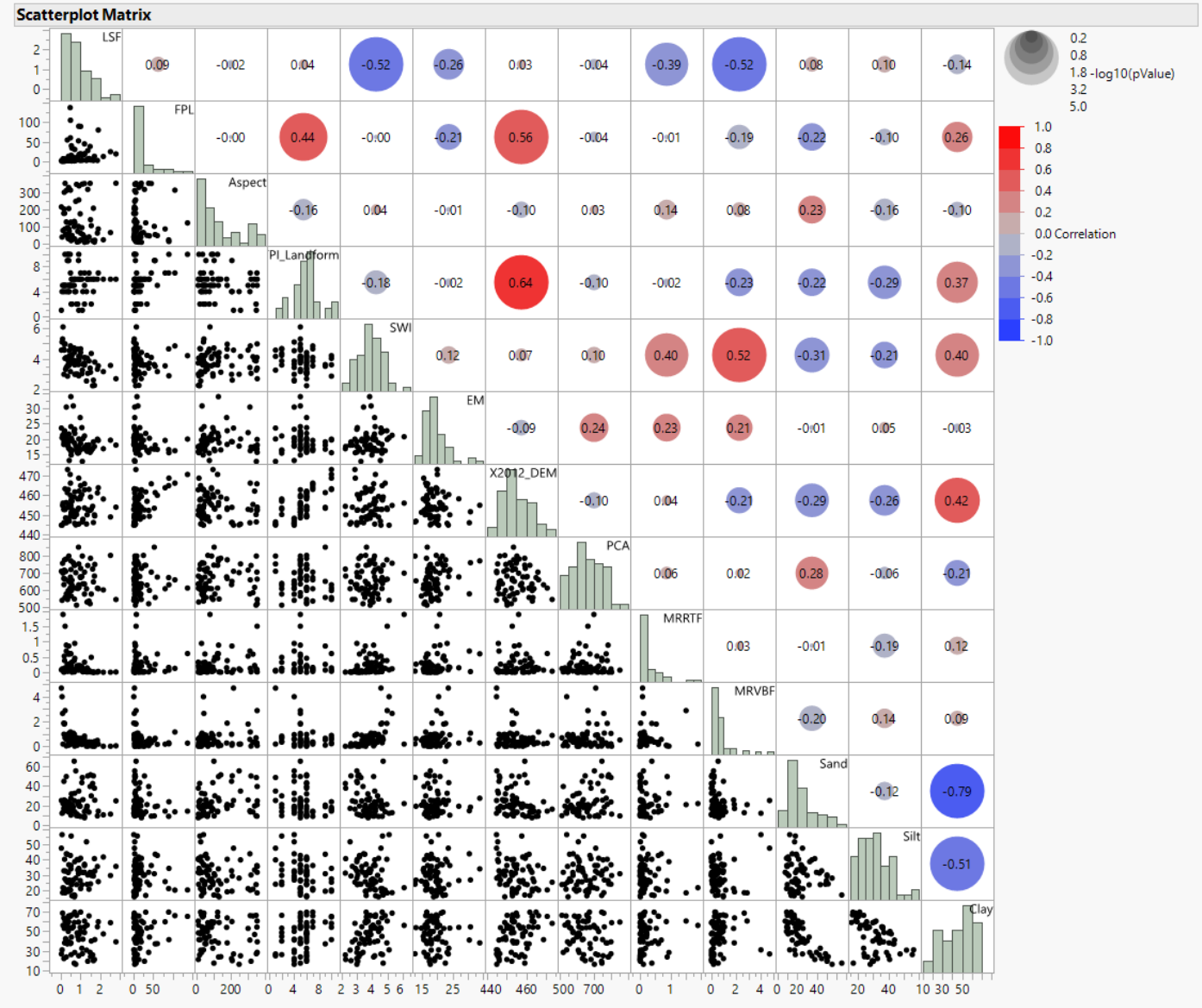

Fig. 21 Correlation matrix of cm and p-values of measured sand, silt, clay at 70-100 cm and environmental covariates.

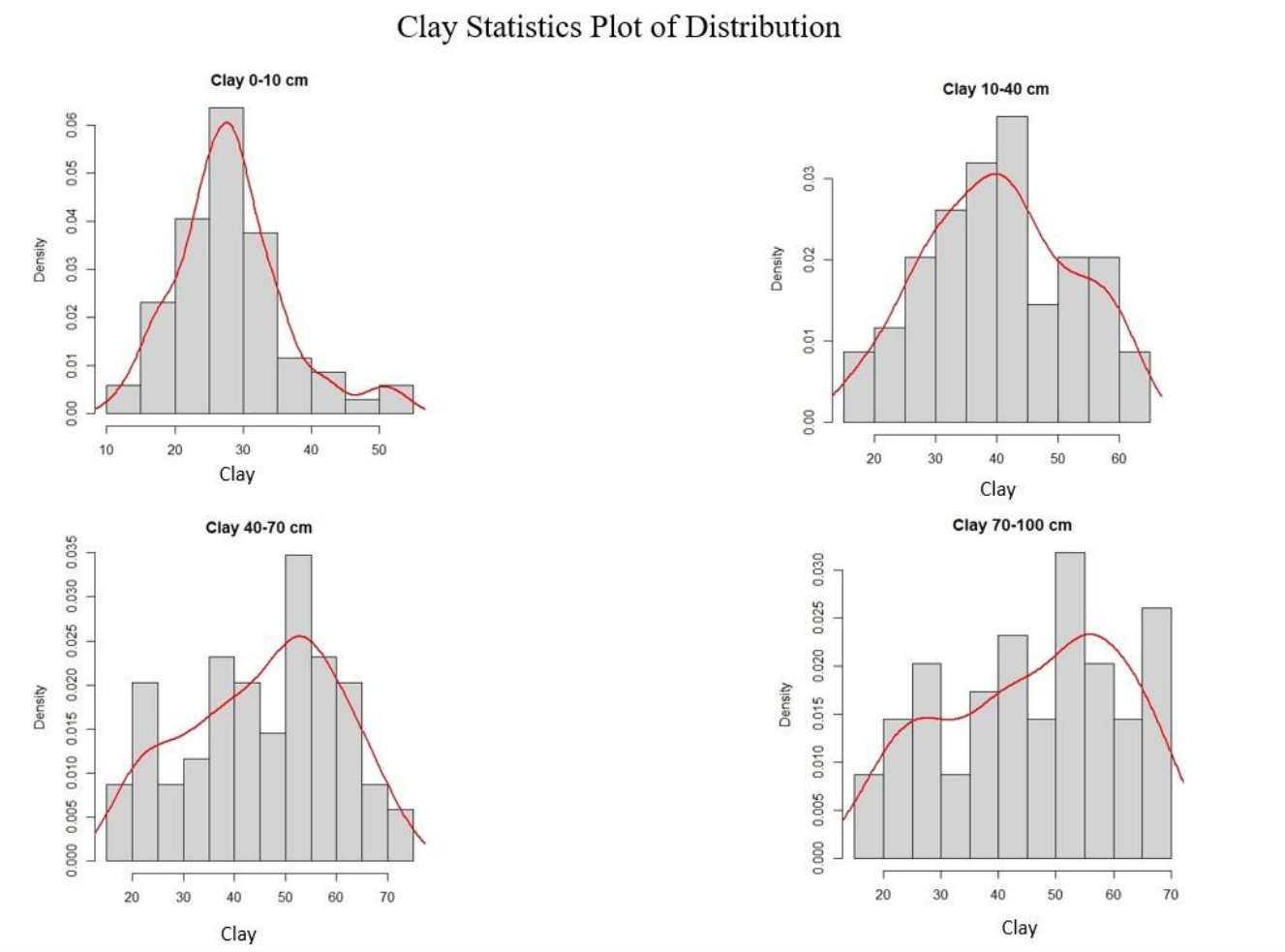

Fig 22. Clay distribution by depth.

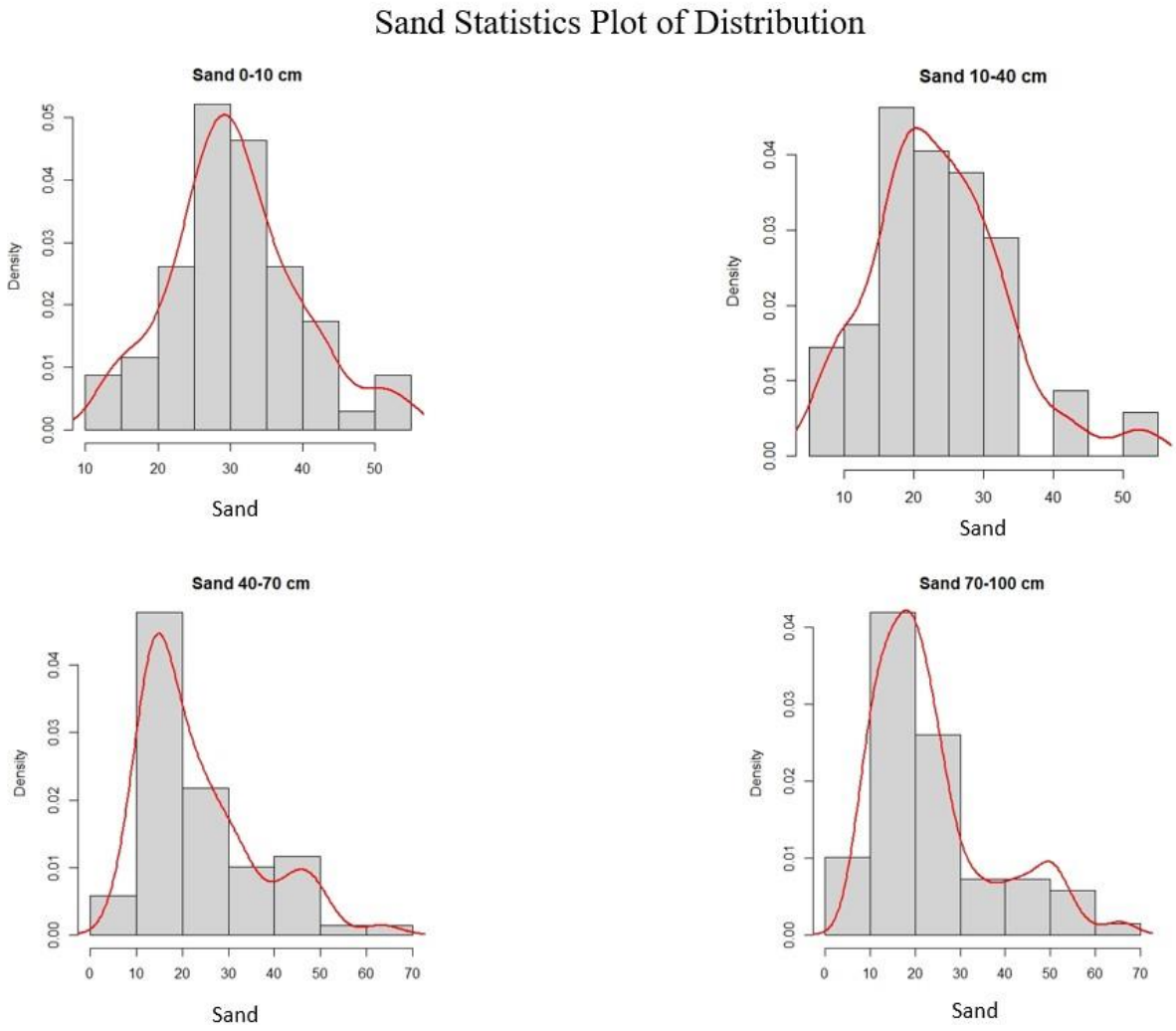

Fig. 23. Sand distribution by depth.

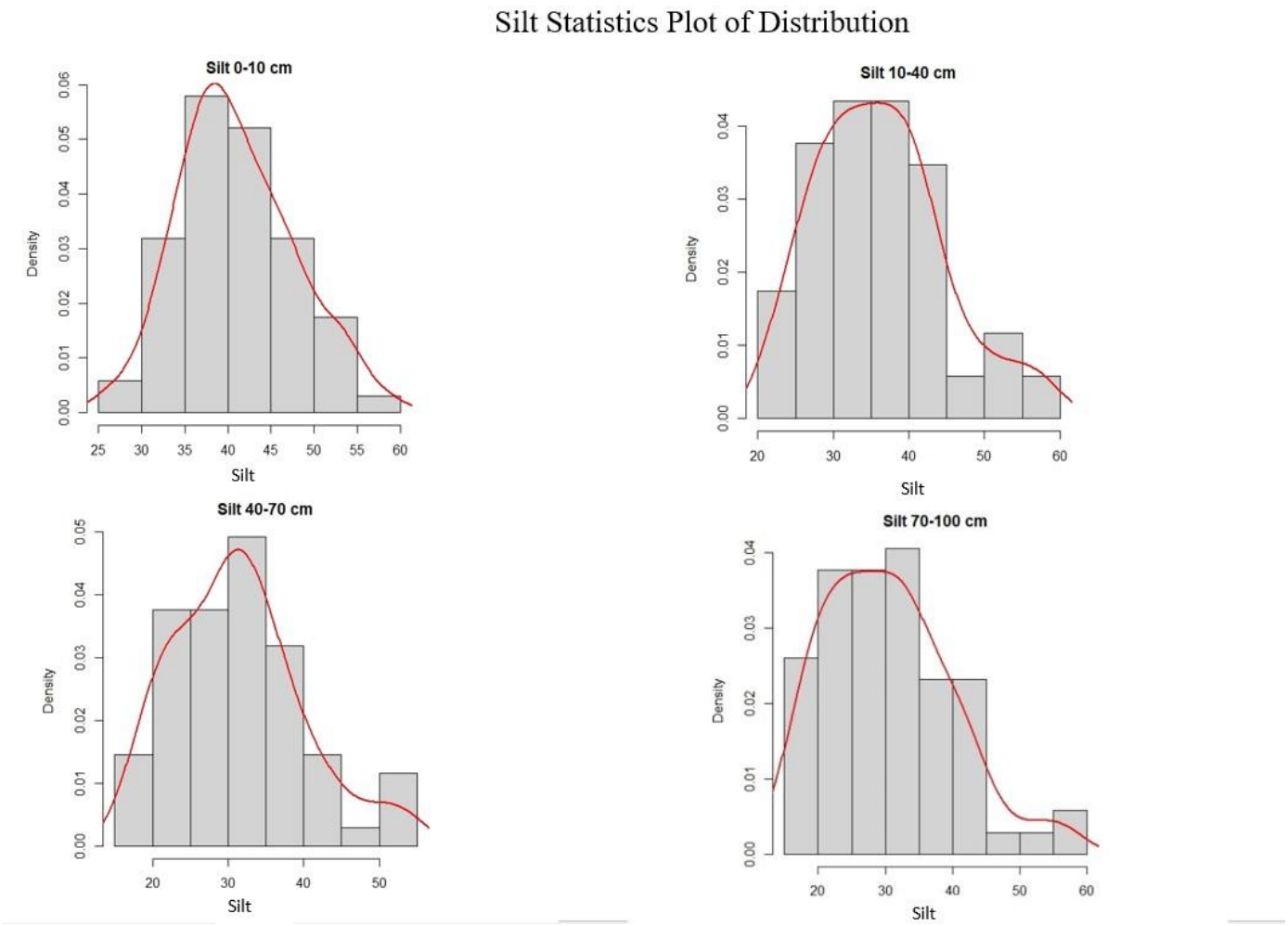

Fig. 24. Silt distribution by depth.