2013

Solving for y: digital soil mapping using statistical models and improved models of land surface geometry

Stephen M. Roecker
West Virginia University

Solving for y:

digital soil mapping using statistical models and improved models of land surface geometry

Stephen M. Roecker

Thesis submitted to the
Davis College of Agriculture, Natural Resources & Design
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Science
In
Plant and Soil Sciences

James A. Thompson, Ph.D., Chair
E. James Harner, Ph.D.
Jeffrey G. Skousen, Ph.D.

Division of Plant and Soil Sciences

Morgantown, West Virginia
2012

Keywords: digital soil mapping, soil survey, geomorphometry, digital terrain analysis, statistical modeling

Abstract

Solving for y: digital soil mapping using statistical models and improved estimates of land surface geometry

STEPHEN M. ROECKER

Digital soil mapping (DSM) is a rapidly growing area of soil research that has great potential for enhancing soil survey activities and advancing knowledge of soil-landscape relationships. To date many successful studies have shown that geographic datasets can be used to model soil spatial variation. This thesis addresses two issues relevant to DSM, scale effects on digital elevation models, and predicting soil properties. The first issue examined was the effect of spatial extent on the calculation of geometric land surface parameters (LSP) (e.g. slope gradient). This is a significant issue as they represent some of the most common predictors used in DSM. To examine this issue two case studies were designed. The first evaluated the systematic effects of varying both grid and neighborhood size on LSP, while the second examined how the correlation between soil and LSP vary with grid and neighborhood size. Results of the first case study demonstrate that finer grid sizes were more sensitive to the scale of LSP calculation than larger grid sizes. While the magnitude of effect was diminished when comparing a high relief landscape to a low relief landscape, the shape and location of the effect was similar. Results of the second case study showed that the correlation between soil properties and slope curvatures were similarly optimized when varying the spatial extent, but that the effect was more sensitive to grid size than neighborhood size. Slope gradient also showed significant correlations with some of the soil properties, but was not sensitive to changes in grid or neighborhood size.

The second study attempted to predict numerous physical and chemical soil properties for several depth intervals (0-15, 15-60, 60-100, and 100-150-centimeters), using generalized linear models (GLM) and geographic datasets. The area examined was the Upper Gauley Watershed on the Monongahela National Forest, which covers approximately 82,500 acres (33,400 hectares). This watershed represents a complex landscape with contrasting geologic strata, deciduous and coniferous forests, and steep slopes. Given this landscape diversity it was still possible to fit GLM which explained on average 38 percent of the adjusted deviance for rock fragment content, and exchangeable calcium and magnesium, and phosphorus. Some of the most commonly selected environmental predictors were slope curvatures, lithology types, and relative slope position indices. This seems to validate the prominence of these variables in theoretical soil-landscape models. Had the correlation between the soil properties and slope curvatures not been optimized by varying the spatial extent, it is likely that another less suitable LSP would have been selected.

Acknowledgments

First and foremost I would like to mention those individuals who helped do the dirty work. Jared Wilmoth helped dig and describe almost ever soil profile collected for this study during the summer of 2006. This required Jared to put up with rain, more rain, lighting, bugs, hiking, tent camping, and long days. Without his help I would never have been able to collect my field data in one summer. Through it all Jared kept his famous sense of humor. However, Jared had his limits, he would not work weekends. That pleasure fell to my girlfriend and now wife, Jennifer Hendricks. She sportingly agreed to help me backpack into those sites that were too far from a road to sample in a single day.

After the summer of 2006 came to a close, the fun part began. During which time Tara Matheny-Helmick served as my laboratory assistant. She did much of the tedious work processing the soil samples for texture, extractable cations, and pH. She showed remarkable attention to detail and also displayed a good sense of humor. Without her assistance I wouldn't have been able to devote sufficient time to my studies. Equally tedious work was delegated to Randy Riddle, who entered many of my soil profile descriptions into the National Soil Information System (NASIS). He almost volunteered for Jared's position, but had the better sense not too.

This study like many others relied on the previous hard work of others. Some notable contributions to this thesis are as follows. Cara Sponaugle, a previous WVU student collected a soil dataset from an adjacent watershed on the Monongahela National Forest which included bulk density, which is necessary to convert soil chemical measurements to a weight basis. With her dataset I was able to develop a pedotransfer function to estimate bulk density for my soil horizons. Soil bulk density is particularly difficult data to collect, and I appreciate her sharing her dataset with me. Dylan Beaudette developed an R package for managing and analyzing soil data which vastly simplified certain aspects of my data analysis. Dylan has always been willing to share his thoughts and code with me, which has helped teach me a thing or two. Lastly, I feel obliged to thank the many developers of the R statistical environment, and SAGA and GRASS geographic information systems. These free software programs were an invaluable resource those synergy and extensive functionality made it possible to iteratively rerun many of my analyses. Thankfully the developers of these programs selflessly distribute their software for the benefit of other scientists.

Unfortunately science is not a cheap endeavor, and would not be possible without financial assistance. Therefore for this my supreme thanks goes to the USDA-NRCS National Soil Survey Center Geospatial Research Unit (GRU) at WVU. Luckily their interested matched mine. Their funding allowed me to learn and contribute to issues concerning the application of new technology to Soil Survey.

Last but not least my thanks goes to my committee members. As expected Dr. Skousen provided his unique insight and fresh perspective which helped me more clearly communicate my ideas. This can be invaluable when a student ventures to deep into the weeds. To Dr. Harner I owe a debt for introducing me to R and instilling in me a

love of statistics; whether he knew it or not. Prior to sitting through his courses I considered statistics to be mathematical wizardry. I still think it is, but now I can honestly say I know how many of the tricks work. Dr. Thompson served as my advisor. As a good advisor should, he pointed me in the right direction whenever I veered astray. However, he also let me take charge and explore avenues that weren't necessarily his first instinct. For this reason, I will always consider him not just a good advisor but a great one.

Contents

List of figures

List of tables

Chapter 1

Soil survey and digital soil mapping

Introduction

Soil is a basic component of ecosystems that performs many ecosystem services including moderating the hydrologic cycle, regulating the major elemental cycles, supporting plant roots, and decomposing wastes (Daily et al., 1997). In order to predict the impact of land use activities, crop responses, waste disposal, storm runoff, and other environmental issues influenced by soil properties, it is necessary to know the variable nature of soil across landscapes. The careful and judicious management of any valuable resource requires that management decisions be based on sound information; lest actions be taken that may otherwise diminish the value of that resource or cause other unintended consequences.

To provide quality soil information for general land use planning, the United States of America (US) has had an active National Cooperative Soil Survey program (NCSS) for over a hundred years. Since the NCSS began, the way in which soil is viewed and the methods available to analyze it have changed considerably, leading to a more complete knowledge of the soil as a collection of natural bodies. Part of the continued success of the NCSS has been attributed to their effort to correlate similar soils across the range of their extent (Arnold, 2006), and their focus on making interpretive soil maps. Soil information generated by the NCSS is utilized by the government and public as an unbiased tool for appraising the productive capacity and quality of a given piece of land. State and county levels of government in particular use soil information as a planning tool, and even as a component in tax assessment and regulatory restrictions. The US Department of Agriculture (USDA) is one of the biggest users of soil information for its farm and conservation programs. This broad application of soil information validates the utility of soil surveys as a primary resource assessment tool.

To make a soil survey, soil scientists delineate similar areas of land (i.e. map units) which contain unique combinations of soils, characterize their soil properties, and infer their uses. In order to segment the soil-landscape continuum into meaningful or natural units, soil scientists use geomorphic features such as “topographic divides, contacts between different rocks or sediment, inflections in slope gradient or shape, and contacts between different landforms of different age, origin, and internal structure” (Wysocki et al., 2011). In addition, vegetation communities are also commonly used to identify soil boundaries. Due to practical considerations there are typically insufficient soil observations to statistically estimate the composition of map units. Therefore, the compositions of map units are inferred from soil scientists’ expert intuition, which they acquire by observing the soil at a number of opportune and purposive locations (Hudson, 1992; McKenzie and Austin, 1993).

Limitations of conventional soil mapping

While conventional soil mapping (CSM) methods have been sufficient in the past, it is currently felt that its methods need to change in order to increase their efficiency and accuracy, and address new soil issues. The philosophical impetus for this change can be broken down into four categories: what, where, how, and how well. The 'what' refers to what do we mean when we say soil. Does soil refer to an individual soil taxonomic unit or a particular soil property (e.g. available water holding capacity or base saturation)? Traditionally, the soil-landscape continuum has been conceptualized using hierarchical classification systems. This conceptual soil segmentation is distinct from that used in delineating map units, in that it based on the vertical distribution of soil properties at a single point, and does not take into account their geographic context (e.g. x,y coordinates, slope gradient, landform, etc…). While classification systems offer a useful scientific language, they also have notable issues, such as "Soil Taxonomy creates classes that are only partially related to landform" (Young and Hammer, 2000), "more direct interpretation can be made from property maps" (McKenzie et al., 2000), "assumption of high covariance of soil attributes" (Gessler et al., 1995), and "weak correlations between mapped classes and some soil properties" (Heuvelink and Webster, 2001). For example, one of the most obvious questions that should be asked, are do the classes correspond with the primary soil property of interest? Presently soil carbon is of great interest due to its role in climate change. However, historically soil carbon has not received significant attention in CSM. No classification system can be designed to adequately address all eventualities. Inevitable all classification systems compress information for the sake of hopefully increasing comprehension. However, now that computers offer an alternative to the storing and sorting of information, it is no longer necessary for soil taxonomic units to serve as the primary geographic unit. Instead it is now possible to map individual soil properties.

Intimately related to the question 'what', is the question 'where'. Due to the discrete geographic model used in CSM, the soil-landscape continuum is abstracted into map units. This approach proceeds under the notion that the variability within map units is less than the variability between map units (Heuvelink and Webster, 2001).The consequence of this discretization is that the internal variability of the soil-landscape continuum is reduced to measures of central tendency (e.g. mean) and dispersion (e.g. variance) for each soil within a map unit. If a map unit is composed of a single soil, this might be sufficient. However, in many cases map units are composed of multiple soils. Thus, estimates of their soil properties are generally displayed as a weighted average of the soil components listed in a given map unit. This makes it difficult to manipulate and combine them with other forms of continuous geographic information for environmental modeling (Zhu, 2006), such as in precision agriculture where the exact location of individual soils is of interest. In some cases such spatial variability can be delineated, but it is considered impractical to do so in CSM.

Even if the issues of 'what' and 'where' could be satisfied, there is also the issue of 'how' we know what we know, and 'how well' we know what we know. As mentioned earlier, map unit delineations and compositions are

estimated by soil scientists' according to their expert intuition. Their intuition is honed by experience and education, and is superior to other scientists in this area of study, but it is still biased and unrepeatable. To be sure, there are standards and guidelines for conducting soil surveys, but there is no explicit rule set or parameters given for 'how' a given map unit was derived. For these reasons Hudson (1992) concluded that CSM is "overly dependent on tacit knowledge", while McKenzie and Ryan (1999) said that "users surveys find it difficult to separate evidence from interpretation" (McKenzie and Ryan, 1999). As there is a significant portion of random variability to landscapes, which cannot be accounted for, it is equally important to know how accurate a soil map is, or how widely the soil properties fluctuate. This requires a formal assessment of soil information's accuracy or uncertainty (i.e. 'how well').

Digital soil mapping

To address the limitations of CSM, soil researchers from around the globe have experimented with a variety new computationally intensive methods capable of predicting or interpolating soil information from field and laboratory soil data. As all these new methods share a common theoretical framework, they have been generically referred to as digital soil mapping (DSM) (McBratney et al., 2003). DSM has also been referred to as environmental soil-landscape modeling (Grunwald, 2006), predictive soil mapping (Hewitt et al., 1993; Scull et al., 2005), and environmental correlation (McKenzie and Gallant, 2007). More than simply a digitizing of existing CSM methods, DSM involves the prediction of soil information by computer models. The degree of computation may vary, but at a minimum DSM generally formalizes soil scientists' expert knowledge into a rule-based framework for spatial prediction, and provides an estimate of the uncertainty of its predictions (McBratney et al., 2002; MacMillan, 2010).

Spatial prediction is typically achieved by one of two means, referred to as the spatial or *clorpt* approaches respectively, which in some circumstances maybe combined. The first approach interpolates predictions to new locations as a function of their distance from neighboring observations, by modeling the spatial dependence between observations with a variogram (e.g. geostatistics). This is an effective technique were the soil exhibits spatial correlation, is sampled at distances closer than the average range of spatial dependence, and can provide interpretable spatial statistics. However in most cases the soil varies so greatly over short distances, that this approach is generally considered impractical for mapping at small scales (i.e. large areas). The second approach derives predictions of soil properties and/or types as a function of their relationship with environmental predictors, with rule-based (heuristic) or statistical models. This is an efficient technique when the environmental predictors are more easily attainable than the soil observations, and have a strong physical connection to the soil characteristics of interest (Gessler et al., 1995). The popularity the second approach stems from its strong theoretical framework, which is based the state factor model

$$s = f(cl,o,r,p,t,...),$$

(Jenny 1941, 1980). Recently, McBratney et al. (2003) has expanded the original formulation of *clorpt* to include existing soil information (s) and spatial location or distance (n), resulting in the acroynm, *scorpan* (note time (t) has been swapped with age (a)). The incorporation these additional factors, recognize their value for prediction. By utilizing both the spatial and *clorpt* or *scorpan* approaches to DSM, researchers have modeled a variety of soil properties and types, over a range of scales, with varying degrees of accuracy and precision. At this point DSM has reached a point where it is considered by many to be ready for operational mapping (Burrough, 1993; MacMillan 2007; Hengl 2009).

In most cases the *clorpt* or *scorpan* approach has been the preferred method; in part due to the vast range environmental predictors now available, such as digital elevation models (DEM) and satellite imagery. Because of the strong influence of water movement on the development of catenary sequences at the mesoscale, DEM derivatives have been shown to be a good predictor of local soil spatial variation. However, the strength of this relationship though has been found to decrease with depth (Florinsky et al., 2002; Park and Vlek, 2002). As depth increases it is believed that soil variation becomes more strongly influenced by vertical pedogenic processes (Park and Vlek, 2002), though such processes are still influenced by topography.

While DSM has made great progress in predicting soil variation, a number of limitations exist, such as nonlinear processes and spatial dependence. Generally, nonlinearity in DSM has not been thoroughly examined in the literature, but is expected to become an issue when predictions are extrapolated over large areas; which should correspond to an increase in the heterogeneity of the environmental factors. At the level of the catena, Gessler et al. (2000) has suggested that linear models are appropriate as they portray soils smooth transition along a hillslope. However, Park and Vlek (2002) built DSM for 32 soil properties and found only soil moisture, pH, and clay content linearly related to DEM derivatives. Spatial dependence, on the other hand is not accounted for in most DSM, which makes them essentially non-spatial. Instead these models are spatially projected based on their functional relationship to other spatial varying predictors. Regardless of these constraints though, the applicability of DSM techniques have reached a stage where they are being used to create useful digital soil products over substantial areas, as demonstrated by Bui and Moran (2003) and MacMillan et al. (2005).

Mathematical and statistical modeling

In order to predict the spatial distribution of soil properties or classes, various mathematical and statistical models can be used. A comprehensive review of their application in DSM is provided by McBratney et al. (2003). Similar reviews have also taken place in ecological modelling (Guisan and Zimmermann, 2000) and geomorphometry (Hengl and MacMillan, 2009). Hengl and MacMillan (2009) in particular highlight the applicability of different models depending on the available data and question at hand. Regardless of the number of comparisons that have occurred in ecological modelling, Austin et al. (2006) stressed that the most important consideration is not the statistical model employed, but the ecological knowledge and statistical skill of the analyst. Minasny and McBrat-

ney (2007) have likewise concluded that improved spatial prediction of soil properties will result from accumulating better soil data, rather than more sophisticated statistical models. Some specifics of both the most common models are described below, including kriging, fuzzy logic, generalized linear models, and tree-based models.

Geostatistics

Geostatistics is family of spatial interpolation methods based on regionalized variable theory, which predicts values at new locations by modeling the spatial dependence between neighboring observation values as a function of their distance. In essence it predicts values at new location by taking a spatially weighted average from neighboring soil property values. However unlike other forms of spatial interpolation, the spatial weights are estimated objectively with a statistical model, the variogram, rather than by an arbitrary mathematical function. The basic form of the variogram is a plot of the lag or distance between point observations (x) against the variance (y). As the lag increases so does the variance up until some point at which the best estimate of a given soil property is the global mean. In addition to interpolating predictions, kriging is able to estimate the variance at each point, which can be used to judge the spatial accuracy of the interpolation. The basic premise of spatial interpolation is that the closer together two points are the more likely they are related. The most basic geostatistical interpolation method is referred to as ordinary kriging.

Two key assumptions of geostatistics are that the properties of interest are the result of a random process and are stationary. Because soil is the result of deterministic processes though, Webster (2000) remarks that clearly the soil is not random but instead chaotic. But because we cannot determine the difference between those processes that are either random or deterministic, it makes no difference whether or not we model the soil as if it were random. The assumption of stationary implies that the mean and variance are constant throughout the region of interest, and that the variance does not increase with increasing area. This assumption truly has no answer though because there is only one realization of the generating process in a particular region (Webster, 2000). So in geostatistics, it is not the soil that is random and stationary, but the model, which may be one or both. Instead Webster (2000) asserts that the real question is whether or not a stationary model is realistic (given the circumstances), and leads to accurate predictions.

Kriging is a suitable method in presence of spatial dependence. However in many cases, soil properties are the result of deterministic processes. In such cases it is beneficial to model the deterministic component of soil spatial variation as a function of ancillary data (e.g. slope gradient and surface reflectance), and any residual stochastic component by kriging. Variants of kriging which incorporates both deterministic and stochastic components include cokriging and regression kriging. In comparison of other statistical and geostatistical models, Bishop and McBratney (2001) have demonstrated regression kriging to be superior. However, at the landscape scale, when the soil is not sampled at distances closer than the average range of spatial dependence, Scull et al. (2005) found multiple linear regression to be superior to regression kriging.

Fuzzy logic

Fuzzy logic is an alternative to Boolean logic that determines the membership to a given class by either a 0 (no) or 1 (yes). Fuzzy logic deals with the ambiguity of defining the soil-landscape continuum by allowing a soil to have partial membership to more than one class, on a scale between 0 and 1. Unlike geostatistics, fuzzy logic is not truly a statistical model, because "it does not assess the accuracy of its predictions" (Heuvelink and Webster, 2001). The distinction between fuzzy logic and Boolean logic is that fuzzy logic is based on possibility theory, while Boolean logic is based on probability theory. In this way fuzzy logic is a measure of a soils similarity to a class, rather than its chance of belong to it (Zhu, 2006). Zhu (2006) asserts that "soil classification is based on possibility, not probability", and as such fuzzy logic is a more appropriate approach for defining soil classes.

The advantage of fuzzy logic is that it allows for representing the continuous nature of both soils geographic distribution and attribute distinctness. The most prominent application of fuzzy logic in DSM has been the SoLIM (Soil-Landscape Inference Model) model, developed by Zhu and Band (1994), Zhu (1997a,b), and Zhu et al. (1996, 1997). This approach uses the expert knowledge of an experienced soil scientist to formalize the relationship between soil series and ancillary data. The incorporation of soil scientists' expert knowledge though can be seen as both an advantage and disadvantage (Scull et al., 2003). The advantage being that it can explicitly summarize a soil scientist's expert knowledge, which has been accumulated at great expense. The disadvantage is that a soil scientist's expert knowledge is subjective and lacks statistical grounds for inference.

Generalized linear models

One of the most commonly used group of regression and classification models are generalized linear models (GLM), which are a modified form of the classical linear model designed to handle situations in which the linear models main assumptions are not met. Those assumptions being that the response is normally distributed with a constant variance and that the predictors combine additively on the response. Lane (2002) has advocated the use GLM in the soil sciences as opposed to transforming the linear model when these assumptions are not met, such as for binomial (presence/absence) and Poisson (counts) distributions. Transformations are typically used to modify the linear model to handle alternative distributions, but can affect the interpretation of additivity on the transformed scale, where statistics like standard errors and variance ratio values should be used with caution (Webster, 2001). Like linear models, GLM have similar fitting procedures and diagnostics so they can be likewise interpreted.

To modify the classical linear model, GLM allow the response to belong to a wide range of exponential family distributions (e.g. Gaussian, binominal, Poisson, Gamma), and relate the response's mean to the model on a scale where the effects combine additively through the link function. The effect of the link function transforms the model to linearity, and maintains the response's range of values. More simply put, this transforms the model, rather than transforming the data to fit the model's assumptions. A consequence of modifying the linear model requires

that the parameters be estimated iteratively by maximum-likelihood, as opposed to being derived analytically as with least squares. Another consequence eliminates the ability to employ the analysis of variance. Instead the analysis of deviance is used, which is a measure of the difference between the observations and the fitted model, which for Gaussian distributions equates to the residual sum of squares.

Aside from being able to handle multiple distributions, GLM have additional benefits, such as being able to use both categorical and continuous predictor variables. Also as with linear models, they allow interactions between the predictors and polynomial terms, so as to model more complex data structures. To identify such interactions exploratory techniques such as tree-based models (Guisan et al., 2002) and coplots (McKenzie and Jacquier, 1997) may be used. The use of interactions though can increase colinearity within the model at the expense of identifying meaningful relationships between variables (Park and Vlek, 2002).

Tree-based models

Tree-based models or decision trees differ from GLM in that they do not make assumptions about the form of the data. Instead they are often referred to as data driven, whereby the resulting models structure is based off of the data itself, rather than some assumed distribution such Gaussian or other. This can be seen as both an advantage and disadvantage. For example, given a sizable data set trees can easily identify complex data structure. In the absence of a sizable data set other parametric models (Maindonald and Braun, 2007) such as GLM are likely to provide better estimates, given that they make assumptions about the structure of the data.

To understand tree-based models, it is best to discuss how they are grown. The standard method of tree construction develops a set of decision rules using binary partitioning, which repeatedly subdivides the response into two sets of increasingly more homogeneous groups until no further purity within the groups can be gained by splitting them. When plotted these decision rules resemble a tree. During each step of the tree's growth the partition of the response is based upon whatever split amongst the predictors creates the best fit. For continuous responses (regression trees) the splitting criteria used is the residual sum of squares, while for categorical responses (classification trees) there are a choice of three splitting criteria, all of which seek to optimize the proportion of correctly classified observations. After the tree is grown, the final groups or leafs are labeled with the mean (regression trees) or majority (classification trees) response within the leaves. While growing a tree following this procedure can be simply automated, the decision of when to stop its growth requires the subjective intervention of the analyst. Ultimately the process could continue until each observation is correctly classified. While this would accurately describe the given data set, it would over fit the existing data set, and therefore poorly predict new data. The idea is that as the tree grows, less and less reduction in deviance is gained with each split. So the tree's overall accuracy would suffer little if it were pruned to a smaller number of leafs. To determine an optimum stopping point, cross validation is used. This pruning method produces a plot of the number of leafs against the amount of devi-

ance explained. The optimal stopping point is the location on the plot where the slope flattens out, or falls below one standard deviation of the minimum cross validated error.

Because of the automated nature by which trees are grown, they are often useful for exploratory data analysis (Guisan et al., 2002), so as to indicate the relative importance and potential interaction between predictors. Also the results of a tree-based models are easily interpretable if the number of binary splits is small, and allow a mixture of both continuous and categorical predictors. Despite the relative ease with which trees are constructed, Hastie et al. (2009) lists three notable limitations. The first being that trees are inherently unstable due to their data-driven nature of construction. As such any change in the data may produce a different tree. For this reason research in DSM (Park and Vlek, 2002; Scull et al., 2005b) has shown tree-based models to perform less well than parametric models when validated by in independent data set. A second limitation of trees is that for continuous responses they produced do not result in continuous predictions, but rather unrealistic stepped predictions. Still for noisy data sets, McKenzie and Ryan (1999) have suggested that this is not a problem. Lastly, Hastie et al. (2009) cite trees inability to capture additive structure. Hastie et al. (2009) state that it possible for trees to capture such structure with sufficient data, but that tree-based models construction process does not readily exploit such structure within data.

In an effort to overcome tree-based model's limitations, a number of alterations to the construction of trees has been proposed, such as boosting, bagging (Breiman, 1996), and random forests (Breiman, 2001). Each distinct alteration creates a model comprised of multiple trees, generally termed an ensemble, or continuing with the use of tree metaphors a forest. By growing a forest rather than a single tree is it possible to take a majority or weighted vote amongst the trees, thereby increasing the accuracy and decreasing the sensitivity of the model. In boosting, a forest is grown by repeatedly reweighing the misclassified and correctly classified observations in the data set. In bagging, a forest is grown by taking repeated bootstrap samples of the observations in the data set. In random forest, a forest is grown by taking repeated bootstrap samples of the observations and predictors in the dataset. While each of these ensemble methods typically generates better estimates than a single tree, they also come at the expense of their interpretability.

Model assessment

The final step in any modeling process should include an assessment of the uncertainty or error associated with the model. This step evaluates the quality of the model's predictions, and is generally referred to as model assessment in statistics or accuracy assessment in remote sensing. In order to construct a model which is useful for predicting future observations a balance between prediction error and model complexity must be achieved (Hastie et al., 2009). What is desired is a model which not only closely approximates the pattern of the observed data, but also that of the greater population from which it comes. Ryan et al. (2000) lists three reasons why statistical models may have low accuracy: poor correlation between the soil and environmental variables, extreme variation pre-

sent within a local neighborhood (i.e. dominated by nugget variance), observed data represents a small range of population.

Sampling

When developing digital soil maps using statistical models, it is necessary that the soil samples be collected in an explicit and probabilistic fashion. This approach is based on classical sampling theory, where the source of randomness comes from the design rather than the model as in geostatistics (Brus and Gruijter, 1997). Randomness in some fashion is necessary to generate unbiased estimates of the population parameters so as to ensure that the results are repeatable, and not the byproduct of chance. Other important considerations include clearly defining which soil individuals are being sampled, specifying their dimensions, and accurately determining their spatial location. The later consideration is important so as to be able to co-register the site locations with the environmental covariates, which are used for prediction. This is now easily facilitated with the use of global positioning systems (GPS). How to best allocate sample sites across the landscape in an unbiased fashion is a more contentious issue and guidance on this issue for soil survey applications are provided by Webster and Oliver (1990), Domburg et al. (1997), and Brus and de Gruijter (1997). Prior to outlining the approach used here, a brief review of that theory is provided. The most common probabilistic sampling strategies that have been utilized in environmental correlation are the simple-random, systematic, and stratified-random designs.

Simple-random

In simple-random sampling, each site has an equally probable chance of being selected. The effect of which typically results in an uneven geographic distribution of sites, which may under represent certain areas (Webster and Oliver, 1990). This makes simple-random sampling an inefficient strategy because the only alternative to achieve a more complete coverage is to sample more sites. Nevertheless, this sampling strategy has been successful in identifying soil-landscape relationships over large areas. Sample sites are easily allocated within a GIS by defining the area of interest, and using a random number generator for selecting the geographic coordinates.

The literature presents two recent examples of this sampling strategy (Bell et al., 1992, 1994; Howell et al., 2004). Bell et al. (1992, 1994) used simple-random sampling successfully to predict soil drainage class using discriminant analysis, with 305 samples collected over the extent of a USGS quadrangle (14,448 ha) in Pennsylvania. The results of this study found that the soil drainage class was correctly predicted at 81, 74, and 69 percent of the sites, based on the calibration dataset, validation dataset, and soil survey, respectively. Howell et al. (2004) on the other hand compared the efficiency of simple-random sampling with purposive sampling, to predict the presence/absence of soil morphological features, using 97 and 656 samples respectively, collected from an ongoing soil survey over an area of 30,424 ha in the Mojave Desert. Purposive sampling is typically performed in conventional soil survey, whereby soil scientists identify sample sites across the landscape using their professional intuition. This

strategy is designed to establish the soil-landscape relationships, on which the soil scientists base their conceptual models (McKenzie and Austin, 1993). In the Howell et al. (2004) comparison, the models produced by the simple-random samples produced a "much more sensitive, more accurate, and greater range of estimated values" than the models from the subjective samples. The models created from the simple-random samples outperformed the subjective samples by 1 to 18%, averaging 10% improved performance. While no interpretation of these results were provided by Howell et al. (2004), it may be that the soil scientists' purposive samples did a poorer job of covering the range of the environmental covariates. This may lend credence to McKenzie and Ryan (1999) speculation that purposive sampling may introduce bias into a soil scientist's conceptual model. Buol et al. (1997) has also recognized a subconscious bias of soil scientists to sample soils for characterization that have better developed morphological features that are not truly representative. The implications of the Howell et al. (2004) results suggest that in the presence of an adequate range of environmental covariates, some form of probabilistic sampling strategy may be more efficient at developing quantitative estimates of the soil-landscape relationships.

Systematic

Systematic sampling as the name suggests, samples sites at predefined equal intervals. A transect would be a one dimensional example of this strategy, while a grid would be a two dimensional example (Webster and Oliver, 2001). This sampling strategy provides a more even coverage of sample sites across an area, making it more efficient than simple-random sampling, but with some notable limitations. The first of which is that because the sites are equally spaced, there is a chance that some individuals within the population, that do not correspond with the sampling interval maybe overlooked. This bias becomes unacceptably large as the sampling density decreases; therefore it is only a suitable approach over areas smaller than a single field (Webster and Oliver, 1990). A second limitation of this sampling strategy is that because sampling sites are allocated nonrandomly, it cannot truly estimate the sampling error. Still Webster and Oliver (1990) advocate that systematic sampling is much more precise than simple-random sampling.

The application of this sampling strategy is seen quite often in investigations involving only single hillslopes (Odeh et al., 1991; Moore et al., 1993; Thompson et al. 1997; Young and Hammer, 2000; Park et al., 2001; Park and Burt, 2002; Florinsky et al., 2002; Chaplot et al., 2000). Typically multiple transects or a single grid are laid across all the major landforms within the landscape to capture the range of variation present. The sampling intensity of these studies range from 64 samples for 2 ha, to 247 samples for 66 ha, with an average of 194 samples for 31 ha. One benefit of this sampling strategy is that the sampling density along the hillslopes should allow for characterizing soil variation across a range of scales, and be able to identify points of inflection on the landscape. The sampling intensity of these studies over such a small area though makes them impractical for anything other than scientific research or high-intensive land management. Ideally it would be profitable to be able to extend these predictions to the broader landscape, but this must be done with careful consideration of the similarity and distance

from the environmental settings with which the relationships are derived. Thompson et al. (2006) has examined this possibility, by comparing models developed within different portions of the same physiographic region, using a stratified-random design rather than a systematic, but unfortunately found the models different.

Stratified-random

In stratified-random sampling bias is introduced to optimize the coverage or information extracted from the sample sites, by partitioning the randomization within a study area. This strategy is preferable to simple-random sampling if prior knowledge exists that suggests individual segments of the population are different. Thanks to the wide range of environmental covariates present in GIS, such stratifications can now be performed easily and explicitly. Many different forms of this approach of varying complexities have been developed for environmental correlation (Gessler et al., 1995; McKenzie and Ryan, 1999; Hengl et al., 2003; Park et al., 2001; Minasny and McBratney, 2006). All seek to minimize the overall prediction error by spreading the sampling sites in feature space, geographic space or both so as to cover the multivariate distribution of the ancillary data, and minimize spatial dependence within the models residuals.

The first and most simplistic stratified-random sampling strategy for environmental correlation was developed by Gessler et al. (1995). In the Gessler et al. (1995) design, the landscape was stratified into equal areas based on the topographic wetness index (TWI). The TWI was used as the stratifying variable because it represented a quantification of catenary position. To avoid the effects of spatial dependence, which should only provide redundant information, Gessler et al. (1995) computed a variogram of TWI to estimate the range of spatial dependence within the landscape. Sites were then randomly placed within the stratified areas at distances further apart than the range of spatial dependence. To extend the Gessler et al. (1995) design to account for a larger area and wider range of environmental factors, McKenzie et al. (2000) incorporated three pedologically significant stratifying variables, one of which was again TWI. In the event that a sampling density of 1 site per 250 ha can be achieved, McKenzie and Gallant (2007) favored this approach.

Hypothesis, motivation and objectives

The hypothesis of this study was that geographic datasets could serve as *scorpan* factors capable of predicting soil properties with statistical models. While this has been demonstrated in other landscapes, there was motivation to examine digital soil mappings applicability within a large and complex landscape such as West Virginia. Such studies are necessary to verify a given method's soundness, as different landscapes represent different challenges. Also numerous studies show a fondness for modeling soil properties rather than soil taxonomic units, which is contrary to conventional soil mapping. This presents an interesting alternative format for soil information. A related issue is the importance of developing appropriate *scorpan* factors. With high-resolution digital elevation models (DEM) now becoming more readily available, there is need to develop new methods capable of dealing with the

large amount short-range variation present within them, as opposed to simply coarsening their resolution. An existing option is to vary the neighborhood size (i.e. spatial extent) used to calculate geometric land surface parameters, such as slope gradient. Given these issues the specific objectives of the following chapters are as follows.

1. The objective of Chapter 2 was to examine how the correlation between soil properties and land surface parameters change with grid and neighborhood size.
2. The objective of Chapter 3 was to develop digital soil maps for several soil properties and depth intervals for a large watershed.

Chapter 2

Scale effects on land surface geometry and environmental correlation

Abstract

The digital representation of the Earth's surface by land surface parameters (LSP) is largely dependent on the scale at which they are computed. Typically the effects of scale on LSP have only been investigated as a function of digital elevation model (DEM) grid size, rather than the neighborhood size over which they are computed. With high-resolution DEM now becoming more readily available, a multi-scale terrain analysis approach may be a more viable option to filter out the large amount of short-range variation present within them, as opposed to coarsening the resolution of a DEM, and thereby more accurately represent soil-landscape processes. To evaluate this hypothesis, two examples were provided. The first study was designed to evaluate the systematic effects of varying both grid and neighborhood size on LSP computed from LiDAR. In a second study, the objective was to examine how the correlations between soil and LSP vary with grid and neighborhood size, so as to provide an empirical measure of what grid and neighborhood size may be most appropriate. Results of the first case study demonstrate that finer grid sizes were more sensitive to the scale of LSP calculation than larger grid sizes. While the magnitude of effect was diminished when comparing a high relief landscape to a low relief landscape, the shape and location of the effect was similar. Results of the second case study showed that the correlation between soil properties and slope curvatures were similarly optimized when varying the spatial extent, but that the effect was more sensitive to grid size than neighborhood size. Slope gradient also showed significant correlations with some of the soil properties, but was not sensitive to changes in grid or neighborhood size.

Introduction

The utilization of digital elevation models (DEM) has proven to be invaluable to recent efforts in digital soil mapping (DSM). According to a survey of the literature by McBratney et al. (2003), DEM were by far the most heavily used form of ancillary data. The popularity of this data source stems from its simple data structure, widespread availability, and most importantly due to the strong influence of topography on landscape scale processes that influence soil variability. From a DEM, multiple land surface parameters (LSP)(also known as terrain attributes) and objects can be extracted, such as slope gradient, slope curvature, solar radiation, catchment area, and compound topographic parameters. The significance of the most common LSP and objects to pedogenesis are provided by Schaetzl and Anderson (2005). These digital representations of the Earth's surface have proven useful in explaining a substantial portion of soil variation with (geo)statistical models. In hopes of explaining more soil variation with LSP, research has sought to improve their digital representations by evaluating the numerous methods

for generating DEM (Chaplot et al., 2006), the algorithms used to derive LSP from them (Florinsky, 1998; Schmidt et al., 2003), and the scale effects due to different grid and neighborhood sizes (Thompson et al. 2001; Smith et al., 2006). Alternative approaches have examined the use of Monte Carlo simulation (Holmes et al., 2000) and wavelets (Gallant and Hutchinson 1997). An entirely different approach has sought to scale up the area (i.e., support) over which soil observations are made, so as to quantify a more representative area of the land surface from which their environmental correlation with LSP are made (O'Connell et al., 2000).

DEM resolution

The scale or spatial extent of LSP is related to two factors, the grid size or horizontal resolution of the DEM used, and the window or neighborhood size over which they are calculated. The importance of grid size can easily be observed by viewing LSP derived over varying grid sizes. It can be seen that as the grid size of the DEM used changes, so does the spatial pattern and detail with which the landscape is represented. This in turn affects the model structure and goodness of fit of soil-landscape models. Because the relationships quantified by soil-landscape models are dependent on the scale at which they are derived, they are not applicable at other scales. This is a minor concern as it takes little effort to develop another model if new elevation data become available. What may be a serious issue for soil-landscape modeling is that while an environmental correlation may be apparent between soil attributes and LSP at one grid size, it maybe unrecognizable at another if present, which is due to the contrasting spatial and temporal scales over which soil-landscape processes operate.

To determine the appropriate grid size for predictive modeling, two approaches exist. The first was developed by Florinsky and Kuryakova (2000), who suggested using a grid size over which a range of soil attributes and LSP exhibit the strongest environmental correlation and occupy a smooth portion within a plot of correlation coefficients versus grid size. By choosing a single grid size within a smooth portion of a correlation plot, Florinsky and Kuryakova (2000) assert that one can avoid heterogeneous spatial variability. While this approach is effective at identifying an optimal grid size, Hengl (2006) points out that it is time consuming process to test all possible grid size combinations, and that no one grid size may be optimal for all LSP. Therefore Hengl (2006) has summarized a list of metrics that can be used to determine an appropriate grid size for representing a given map theme within the logical constraints of the spatial data. Hengl (2006) bases his selection of the appropriate grid size on cartographical and statistical concepts such as: scale, computer processing power, GPS positional accuracy, size of delineations, inspection density, spatial autocorrelation structure and terrain complexity.

In similar studies, many other researchers (Gessler, 1996; Chaplot et al., 2000; Gessler et al., 2000; Wilson et al., 2000; Thompson et al., 2001; Bishop and Minasny, 2006) have examined the effect of DEM resolution on the predictive potential of LSP for soil-landscape modeling. Many have also examined how the LSP themselves change with grid size, so as to better understand how their fundamental representation changes. The general conclusions of these studies have been similar. Increasing the grid size of a DEM produces a smoother landscape with lower

slope gradients on steeper slopes, higher slope gradients on gentler slopes, a narrower range of curvatures, shorter flow-path lengths, smaller values of flow accumulation in lower landscape positions, and greater values of flow accumulation in higher landscape positions. The effect of reduced vertical precision in a DEM results in a less smoothly defined landscape, with more abrupt transitions between slope gradient and curvatures (Thompson et al., 2001). The predictive potential of LSP in these studies have tended to favor finer grid sizes. To summarize, it has been shown that increasing the grid size of a DEM smooths the landscape to the point where key surface features are either lost or distorted beyond a point where they no longer represent the environmental gradients affecting soil spatial variation.

Perhaps the most important assertion that has been made by similar studies is that no perfect DEM resolution exists (Claessens et al, 2005; Hengl, 2006). This is contrary to the common belief that more detailed DEM produce more accurate soil-landscape models. Studies have shown that there are a range of scales over which the soil-landscape relationships exist (Gessler, 1996; Chaplot et al., 2000; Gessler et al., 2000; Florinsky and Kuryakova, 2000; Thompson et al., 2001; Smith et al., 2006). For this reason Thompson et al. (2001) suggested that higher spatial resolution DEM may not be necessary to create useful soil-landscape models. In conjunction with this, Smith et al. (2006) also noted that the large expense of high resolution DEM make them uneconomical for widespread use, considering that high resolution DEM do not always produce the best models. While DEM resolutions of 10-30 meters may prove adequate for most soil-landscape modeling purposes, other studies have found them inadequate in areas of complex terrain and low relief (Wilson et al., 2000; Bishop and Minasny, 2006).

Window or neighborhood size

The second factor to consider in the calculation of LSP from a DEM is the window or neighborhood size. Traditionally the window size used in digital terrain analysis programs such as ESRI's popular ArcGIS Spatial Analyst, is the 3x3 moving window. This method derives the value of LSP at any given point by calculating the first or second-order derivatives based on the elevation values from the surrounding eight cells, a 3x3 window. The neighborhood size by contrast is typically used to refer to the physical distance (i.e. meters) over which measurement is made. By defining the window or neighborhood size in this fashion, the area over which the LSP are calculated is inherently based on the size of the grid cells; because as the grid size changes, so does the distance over which the LSP are calculated. For instance, with a 10-meter grid cell, the distance to the center of its adjacent grid cells is ten meters (except for the diagonal distances, which will be slightly bigger for square windows); whereas for a 30-meter grid cell that distance will be 30-meters. So instead of basing the neighborhood size on the scale of the soil-landscape processes being modeled, the LSP are calculated over an artificial distance (Classen et al., 2005).

The significance of using a fixed window size becomes more readily apparent when deriving LSP from a high-resolution DEM (Shi et al., 2007). This is because when the grid size of a DEM falls below a certain threshold, its fixed neighborhood size is measured over smaller distances than might otherwise be considered representative in

the field. Typically in the field a great deal of short-range variation in topography can be observed, which is purposely ignored by surveyors when measuring slope gradient, in favor of capturing the more general character of the land surface (e.g. signal versus noise). This short-range variation is also inherent in high-resolution DEM derived from remotely sensed devices, which can either be small land surface objects or random noise. To reduce this component of variation within a DEM derived from LiDAR (light detection and ranging), MacMillan et al. (2003) have found it beneficial to resample their 3-meter DEM to 10 meters, and use a series of mean filters (5x5, 5x5, and 7x7) to smooth it before proceeding with the automated analysis and classification of landforms.

Because the processes that affect soil variation are multi-scaled (Lin et al., 2005; Yemefack et al., 2005), it seems reasonable to assume that the effect of LSP, which represent said processes are also multi-scaled and will have an optimal range of spatial extents. An alternative to coarsening the resolution of a DEM to match that of the soil-landscape processes being mimicked by LSP is to expand the neighborhood size over which the LSP are calculated from a DEM. This alternative approach was first suggested by Wood (1996). By increasing the distance over which the LSP are estimated this approach captures more general trends and has the effect of filtering out short-range variation, creating a smoother representation of the landscape. An additional consequence is that the topographic meaning of a particular landscape position can change (i.e., concave becomes convex), if the neighborhood size is larger than a given landscape feature. However, because the grid size is not altered it has a similar effect as coarsening the resolution of a DEM, while preserving more spatial detail. The application of this approach has been demonstrated in a few instances where some have shown the environmental correlation of soil attributes and LSP to vary with neighborhood size (Park et al., 2001; Schmidt and Hewitt, 2004; Smith et al., 2006). In a recent study by Smith et al. (2006), the effect of DEM resolution, neighborhood size, and knowledge implementation on predicting soil series was examined for different landscape positions using the Soil Land Inference Model (SoLIM). Three different experimental designs were carried out in an attempt to isolate the influence of each of the previously mentioned factors. The results of the experiments were similar. First, it was shown that there is a range of scales over which soil series are best predicted. Second, the optimal scale differed between landscape positions. Third, finer DEM resolutions are more sensitive to the choice of neighborhood size. Lastly, the digital scale of knowledge implementation affected the accuracy of the resulting models.

Research hypothesis

With high-resolution DEM now becoming more readily available, I hypothesize that varying the neighborhood size used to calculate LSP maybe a more suitable option to filter out the large amount short-range variation present within them, as opposed to coarsening their resolution. To evaluate this hypothesis, two experiments were performed. The first experiment was designed to evaluate the systematic effects of varying both grid and neighborhood size on land surface parameters computed from LiDAR. In a second experiment was designed to examine how the correlations between soil and LSP vary with grid and neighborhood size.

Methods

Case study 1: systematic effects of varying grid and neighborhood size on land surface geometry

Study area

In this experiment, two separate landscapes were chosen within the West Virginia counties of Gilmer and Jefferson (Figure 2.1 and 2.2). The DEM for each landscape had a grid size of 1-meter, were derived from LiDAR (light detection and ranging) using triangular irregular network (TIN) interpolation, and encompassed approximately 600 hectares. The area used in Gilmer County was the ¼ NW ¼ SE portion of the Aurburn USGS quadrangle; while in Jefferson County the ¼ SE ¼ SW portion of the Shepherdstown USGS quadrangle was chosen.

The location of Gilmer County falls within the Central Allegheny Plateau Major Land Resource Area (MLRA, 126) (USDA, 2006). This region covers the northwestern portion of West Virginia. It is characterized by the plateau's dissected topography, which is dominated by its steep hillslopes, and narrow valleys and ridgetops (Figure 2.1). The geology at this specific site is composed of Permian and Pennsylvanian aged sandstone from the Dunkard group overlying the Monongahela group.

The location of Jefferson County falls within the Northern Appalachian Ridges and Valleys Major Land Resource Area (MLRA, 147) (USDA, 2006). This region covers the northeastern portion of West Virginia. It is characterized by the parallel northeasterly running ridges and valleys, which were formed by the collision of the continental crusts. The topography is characterized by strongly sloping ridges and gently sloping valleys (Figure 2.1). The geology at this specific site is composed of Cambrian aged limestone from the Conococheaque formation.

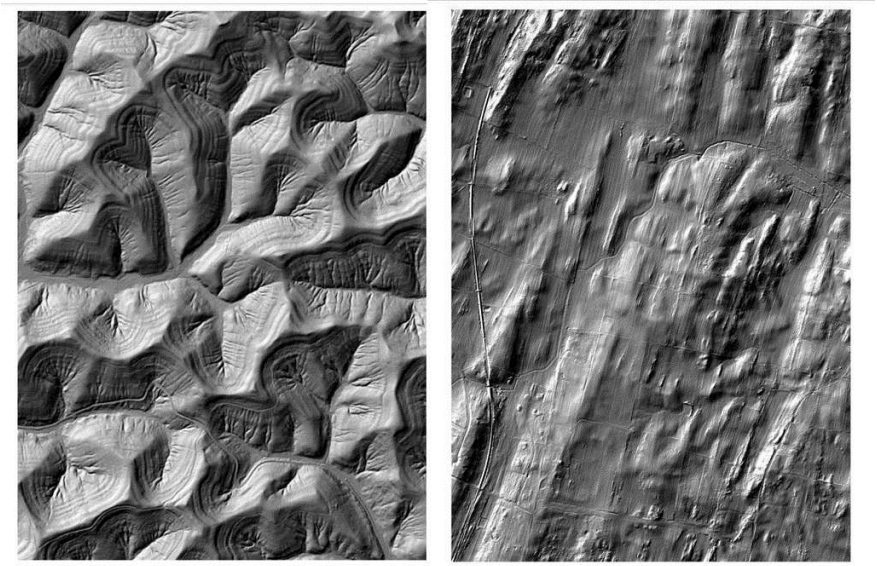

Figure 2.1: Hillshades of landscapes from Case Study 1 (left: Gilmer County; right: Jefferson County).

DEM resampling and LSP calculation

A nested DEM resampling sequence was used in order to compare the different grid and neighborhood sizes at corresponding points and overlapping spatial extents (Table 2.1). In this study the neighborhood size refers to the spatial extent of the window size, rather than the distance between the centers of grid cells. The original 1-meter DEM for each study area was resampled to coarser grid sizes using the average aggregation method. From each DEM, the following LSP were calculated: slope gradient, northerness, profile curvature, and tangential curvature. As slope aspect is a circular measure, its values are not suitable for direct comparison. Therefore, slope aspect was transformed to northerness by, $|180 - \text{Aspect}|$. The GRASS module r.param.scale was used to calculate the LSP, which uses the Evans-Young algorithm (Evans 1972; Young, 1978; Pennock et al., 1987) as implemented by Wood (1996). Wood's (1996) implementation generalizes the Evans-Young algorithm so that the coefficients used to calculate the geometric LSP are estimated from all grid cells within a window, rather the nine grid cells as specified in the original algorithm.

Table 2.1: Experimental contrasts for Case Study 1.

Grid size	1	3	9	27
Neighborhood size	***Window size***			
3	3*			
5	5			
7	7			
9	9	3*		
15	15	5		
21	21	7		
27	27	9	3*	
45	45	15	5	
63		21	7	
81		27	9	3*
135		45	15	5
189			21	7
243			27	9

The window sizes that are labeled with an asterisk (∗) serve as the bench mark, from which the comparisons were made.

Comparison of grid and neighborhood size combinations

To evaluate the systematic effects of varying both grid and neighborhood size on the LSP, exploratory graphical analyses were made. The first approach focused on visually comparing changes in the LSP spatial distribution, as well their non-spatial distribution with box plots. The second approach calculated the goodness of fit between the benchmark window size (i.e 3x3) and expanded window sizes (i.e. 5x5, 7x7, 9x9, 15x15, 21x21, 27x27, and 45x45-cells) for each grid size (i.e. 1,3, 9, and 27-meters). The goodness of fit measures used were the mean difference (MD), root mean square difference (RMSD), and Pearson's correlation coefficient (*r*). Due to the high-resolution of the datasets involved, the number of corresponding points for the finest grid size totaled approximately 5,400,000. Therefore, for the sake of computational efficiency, only the cell centers from the 27-meters DEM were used to evaluate the finer grid sizes, which totaled ~7,400 points.

Figure 2.2: Location of the different study areas (Case Study1: Gilmer and Jefferson; Case Study 2: Upper Gauley).

Case Study 2—Soil and LSP correlations response to neighborhood size

Study area

The study area for the second case study was the Upper Gauley watershed (UGW) within the Monongahela National Forest, which is located on the Appalachian Plateau of southeastern West Virginia (Figure 2.2). Within this area we collected a soil dataset to examine the effect of grid and neighborhood size on the correlations between soil properties and LSP. The DEM came from the Statewide Addressing and Mapping Board (SAMB) and USGS Elevation Conversion Project. This 3-meter elevation dataset was derived using TIN interpolation from mass points and break lines sampled from stereo pair aerial photography.

Figure 2.3: Google Earth overlay of the Upper Gauley watershed and sampling sites.

Soil sampling and analysis

Soil data were sampled from 97 sites within the UGW, using a stratified-random design (Figure 2.3). The stratifying variables used were geology (62% sandstone, 11% shale/sandstone, and 26 % shale), elevation (three quantile classes), and stream power index (five quantile classes). The intersection of these variables created 45 unique strata, within which two random sites were sampled. In order to avoid sampling extraneous features, exclusion rules were used to avoid roads (buffered to 20 meters), streams (buffered to 10 meters), developed areas, and patches of strata smaller than 4,000 square meters (approximately one acre).

At each site a soil pit was excavated to a minimum depth of 140 cm, and described according to standard procedures (Schoeneberger et al., 2002). From each soil horizon, a 300 g soil sample was taken. Each sample was analyzed for particle size (Gee and Bauder, 1983), 1:2 $CaCl_2$ pH, extractable calcium, magnesium, and phosphorus (i.e. Mehlich 1) (Mehlich, 1953), and carbon. Because preliminary analysis showed Ca and Mg to be highly correlated (i.e. 0.89), only their sum was analyzed (i.e. Ca+Mg).The laboratory results from each horizon were aggregated into four depth intervals (0-15, 15-60, 60-100, and 100-150 cm) by taking a weighted average. For the purposes of this study, the thickness of the O horizons was not included. For further details see chapter 3.

Soil and land surface correlation

The procedure outlined by Florinsky and Kuryankova (2000) was used to examine the soil and land surface correlation. This procedure involves plotting the goodness of fit (correlation coefficient, *r*) between the soil and LSP

over a range of spatial extents, and identifying smooth intervals over which the fit was maximized. Smooth portions within the correlation plot are interpreted as representing scales where the correlation is stable, and thus optimal for prediction. In order to utilize soil profiles that were shallower than a given depth internal (e.g. 60-90-cm, rather than 60-100-cm), the correlation matrix was weighted by their thickness.

DEM resampling and LSP calculation

Similar to the first study, the DEM was resampled using the average aggregation method, and the same suite of LSP were calculated using the GRASS module r.param.scale. Rather than evaluate every combination of grid and neighborhood sizes, only one range of grid sizes (3, 6, 9, 15, 27, 45, and 81-meters) and one range of window sizes (3x3, 5x5, 7x7, 9x9, 15x15, 21x21, and 27x27 from the 9-meter DEM) were evaluated (Table 2.2).

Table 2.2: Experimental contrasts from Case Study 2.

Grid size	3	6	9	15	27	45	81
Neighborhood size	***Window size***						
9	3						
24		3					
27			3				
45			5	3			
63			7				
81			9		3		
135			15			3	
189			21				
243			27				3

Results and Discussion

Case Study 1: Systematic effects of varying grid and neighborhood size on LSP

As the neighborhood size increases, the landscape features that are represented by the derived LSP is altered, and short-range variation is filtered out in favor of broader hillslope trends (Figure 2.3). At small neighborhood sizes (e.g., ≤9 m), it is the microtopographic features that are represented by the LSP. As such, there is noticeable short-range variability in LSP values and wide ranges in the distribution of the LSP values (Fig. 2.4). With increasing neighborhood size, the landscape features that are represented by the LSP correspond more closely to recognizable landform elements, such as drainageways, footslopes, backslopes, and shoulders. At intermediate neighborhood sizes (e.g., 9-81 m), the LSP represent a smoother but more connected landscape. This smoothing of the landscape representation is also seen in the boxplots (Fig. 2.4 and 2.5), where the median value remains relatively stable, but the interquartile range decreases and the outliers become less extreme. When larger neighborhood sizes are used (e.g., >81 m), the landscape features depicted by the LSP become oversimplified (Fig. 2.3, 81m).

When the neighborhood size becomes significantly large, the smoothing of the landscape increases and the LSP may misrepresent landform elements because the neighborhood includes DEM data from outside the local landscape (e.g., from across a watershed divide). This distortion or oversimplification of the landscape represented by the LSP appears in the boxplots of slope gradient by a loss in the stability in the median value and continued decrease in the maximum value above a neighborhood size of 81 m (Fig. 2.4).

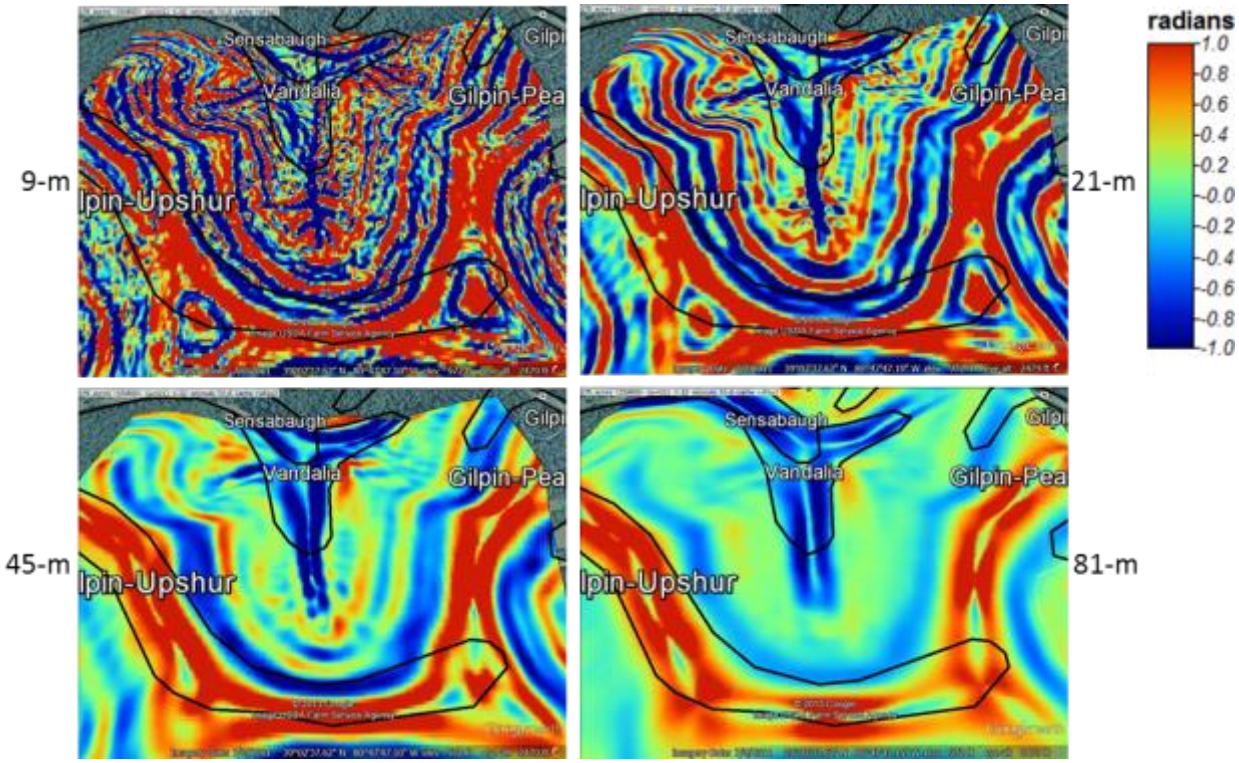

Figure 2.4: Case Study 1. Google Earth overlay of profile curvature and soil lines. Profile curvatures derived from a 3-meters DEM calculated using neighborhood sizes of 9, 21, 45, and 81-meters. Soil lines from SoilWeb (Beaudette and O'Green, 2009a), labeled with the major soil components. The image represents a small catchment from the Auburn QQ quadrangle of Gilmer County, WV.

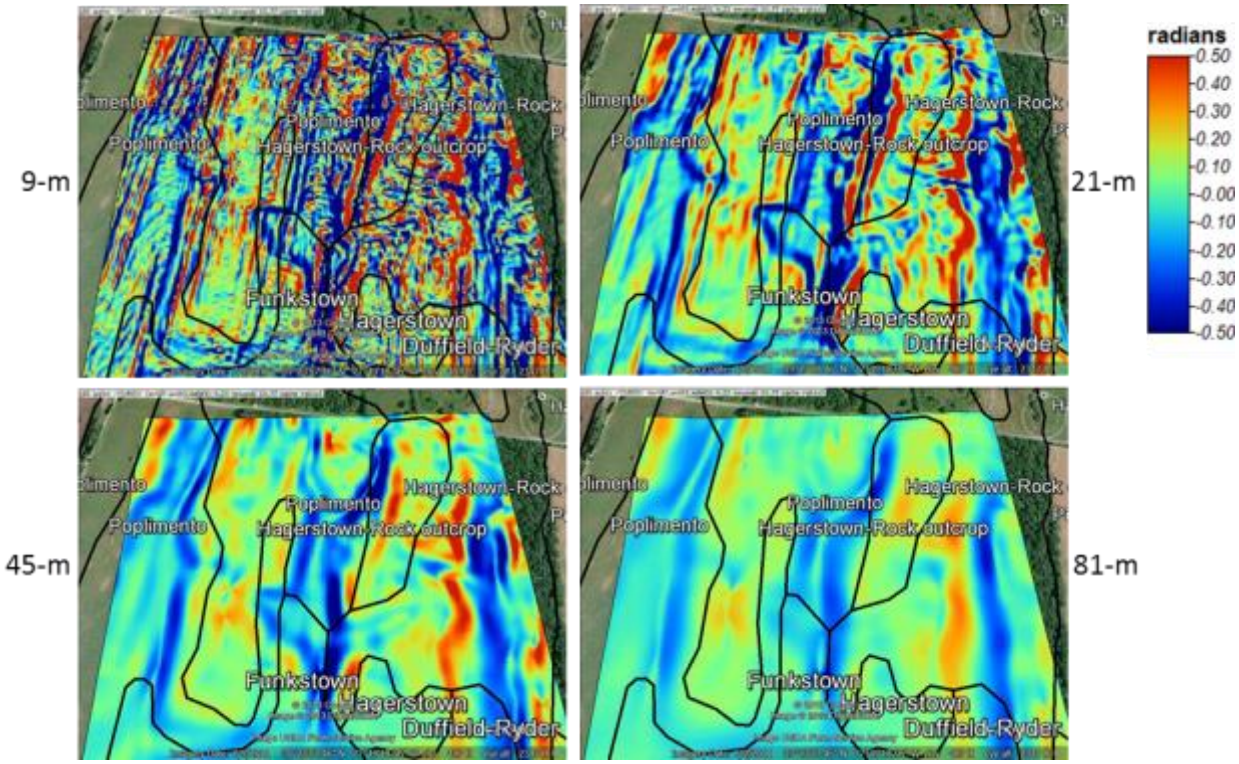

Figure 2.5: Case Study 1. Google Earth overlay of profile curvature and soil lines. Profile curvatures derived from a 3-meters DEM calculated using neighborhood sizes of 9, 21, 45, and 81-meters. Soil lines from SoilWeb (Beaudette and O'Green, 2009a), labeled with the major soil components. The image represents a small catchment from the Shepherdstown QQ quadrangle of Jefferson County, WV.

The effects of changing neighborhood size on the LSP is not the same for the two landscapes examined in this case study. The study area in Gilmer Co., WV, located on the Central Allegheny Plateau, exhibits much steeper slope gradients and more extreme slope curvature values compared to the lower-relief study area from the Northern Appalachian Ridge and Valley of Jefferson Co., WV. Accordingly, the effect of increasing neighborhood size is much less pronounced in the lower relief landscape of Jefferson Co. compared to the higher relief landscape of Gilmer Co. (Fig. 2.4, 2.5, and 2.6). However, in both landscapes the median and maximum slope gradient values begin to drift for neighborhood sizes larger than of 81 m.

While LSP are affected by neighborhood size, the effect of grid size appears to be negligible on their distribution. The same magnitude of decreases in their maximum values, interquartile range, and median value (above a neighborhood size of 81-meters) are seen if slope gradient is calculated using a 1, 3, 9, or 27-meters DEM (Figure 2.6 and 2.7). These and other observations suggest that slope gradient is not sensitive to the effect of spatial extent up until 81-meters. For these landscapes, it appears that a spatial extent of 81m corresponds with a threshold, beyond which all the LSP become increasingly less representative of their land surface shape. Therefore a maxi-

mum grid size of 27-meters, which produces a 3x3 moving window with an extent of 81-meter, appears to be the maximum grid size capable of adequately representing these land surfaces.

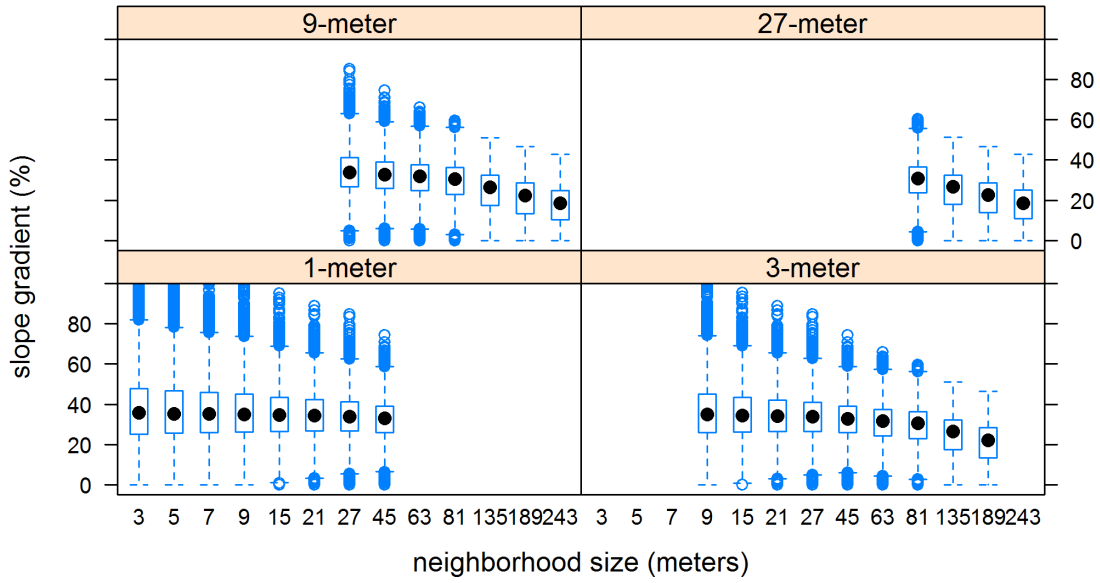

Figure 2.6: Boxplots of slope gradient for Gilmer Co. calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

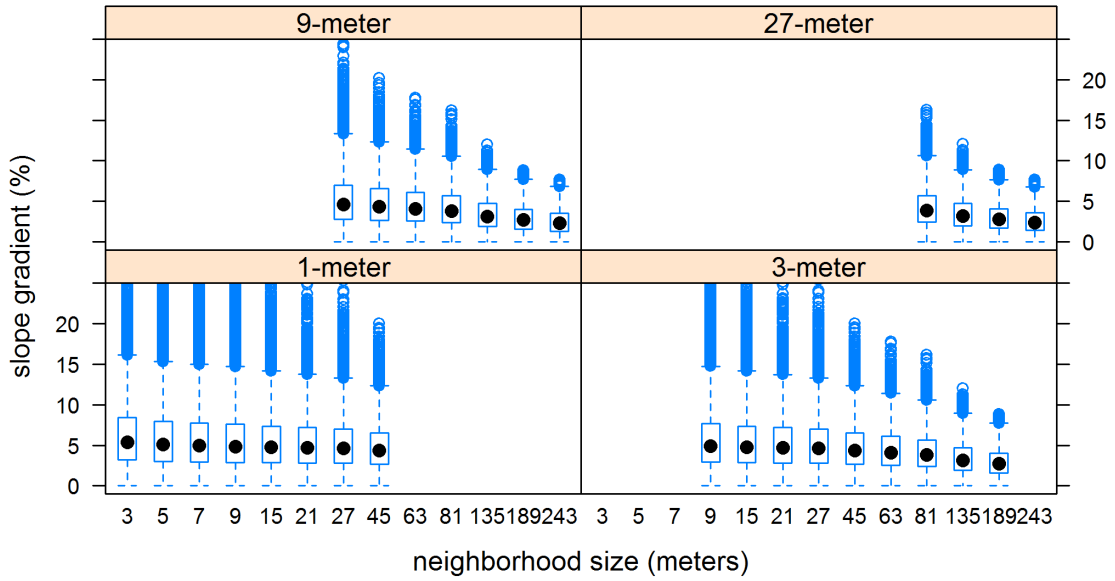

Figure 2.7: Boxplots of slope gradient for Jefferson Co. calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

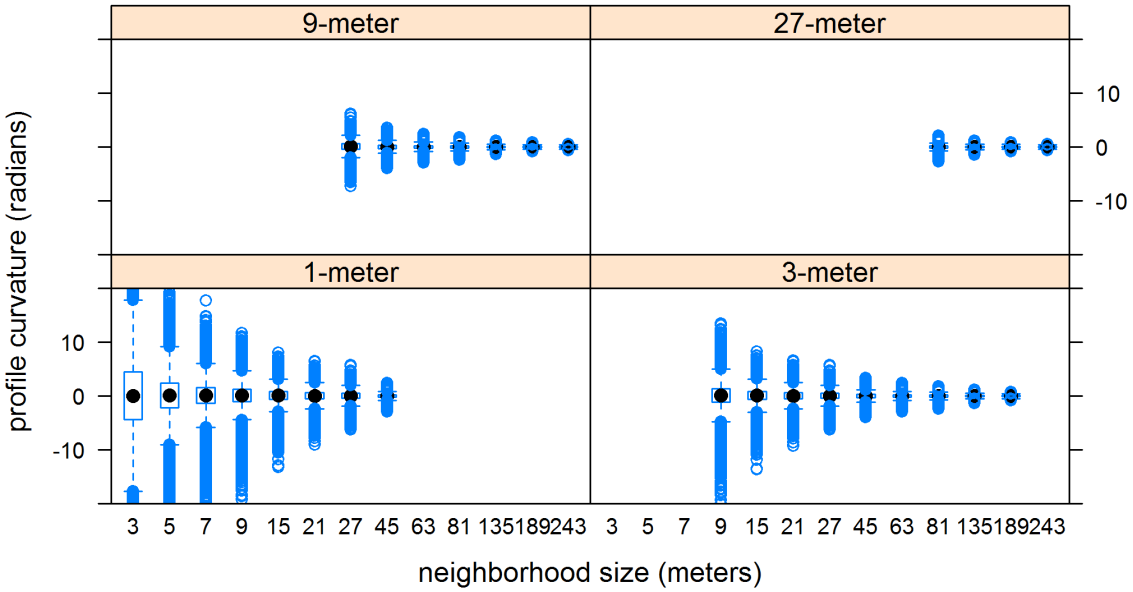

Figure 2.8: Boxplots of profile curvature for Gilmer Co. calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

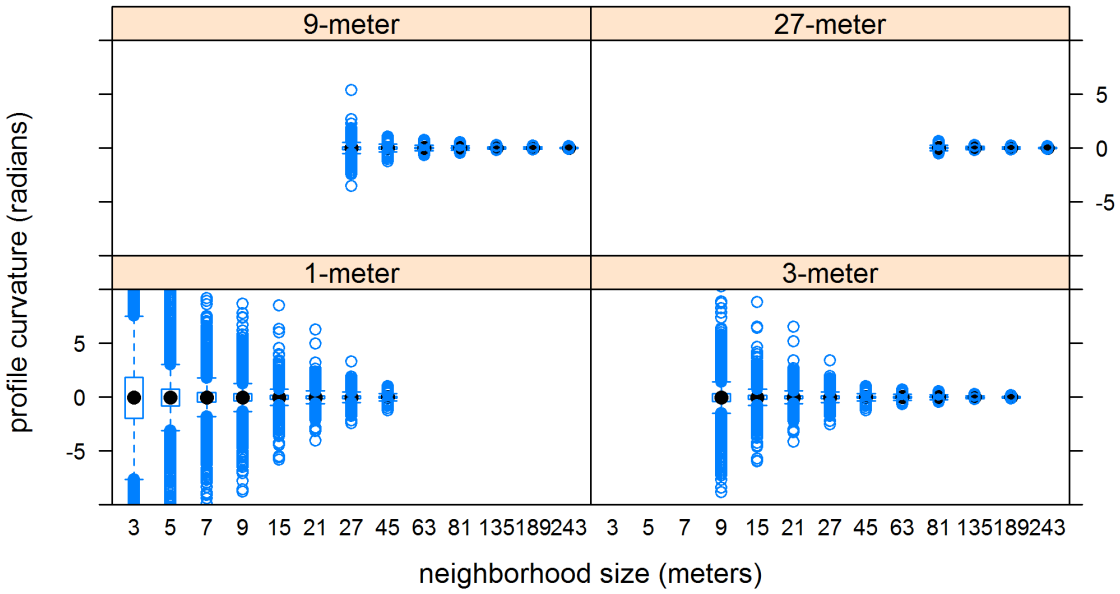

Figure 2.9: Boxplots of profile curvature for Jefferson Co. calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

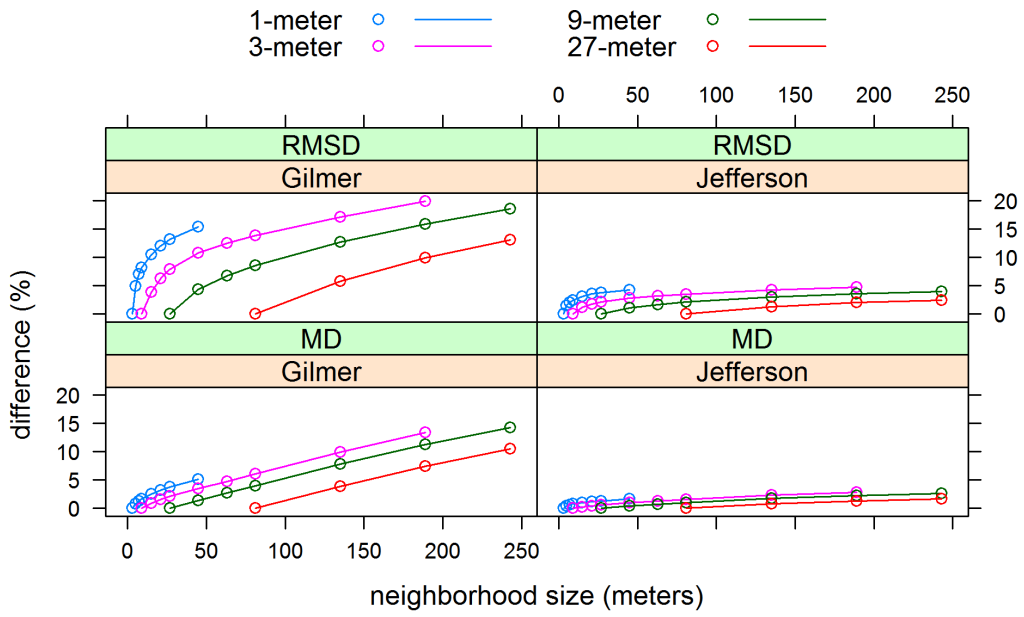

Figure 2.10: Plot of slope gradient mean difference (MD) and root mean square difference (RMSD) goodness of fit measures for Gilmer and Jefferson Co. Slope gradient calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

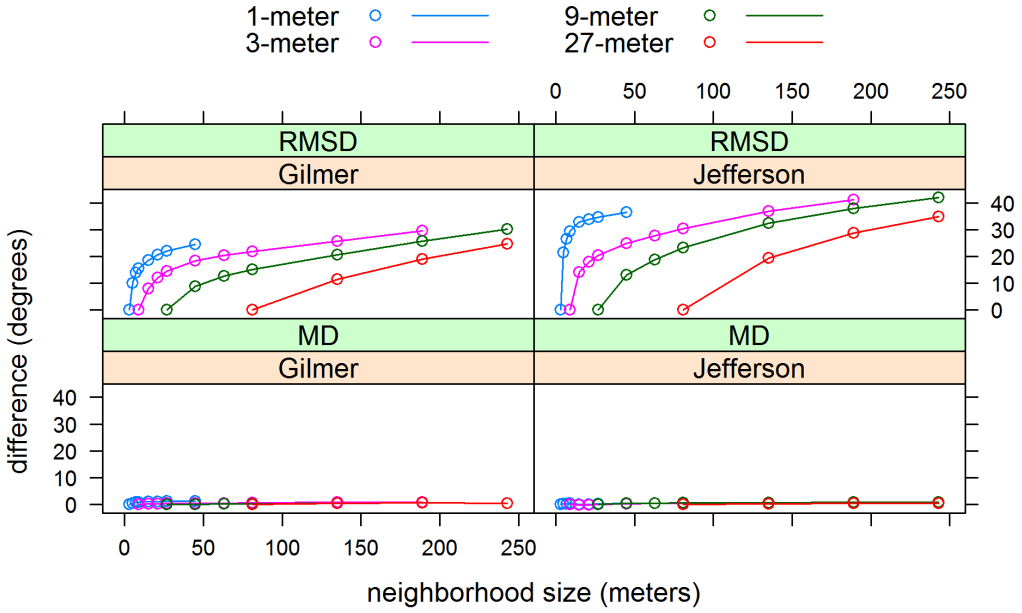

Figure 2.11: Plot of northerness mean difference (MD) and root mean square difference (RMSD) goodness of fit measures for Gilmer and Jefferson Co. Slope gradient calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

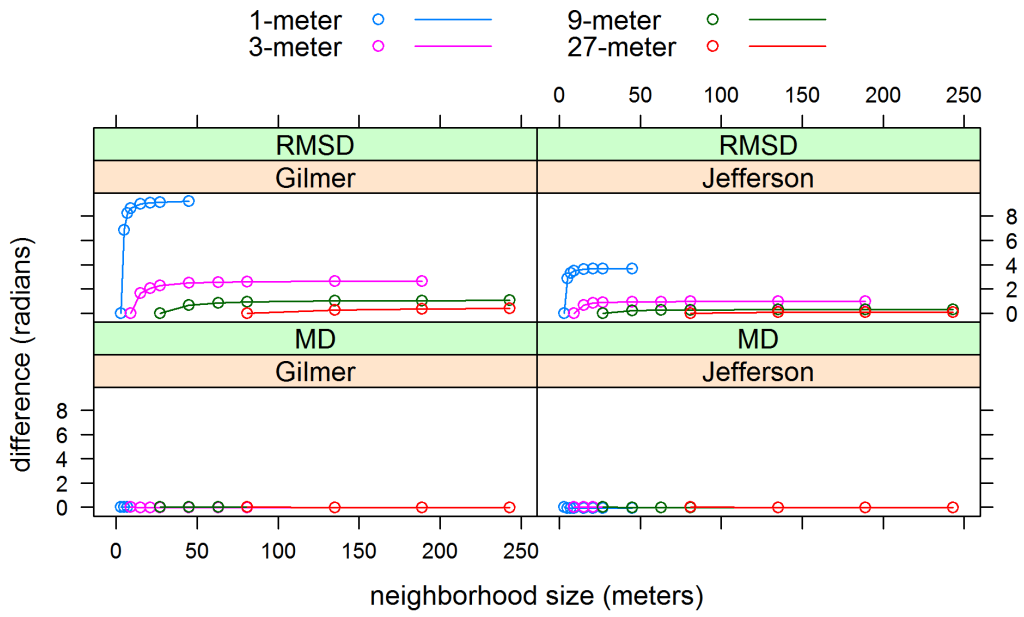

Figure 2.12: Plot of profile curvature mean difference (MD) and root mean square difference (RMSD) goodness of fit measures for Gilmer and Jefferson Co. Slope gradient calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

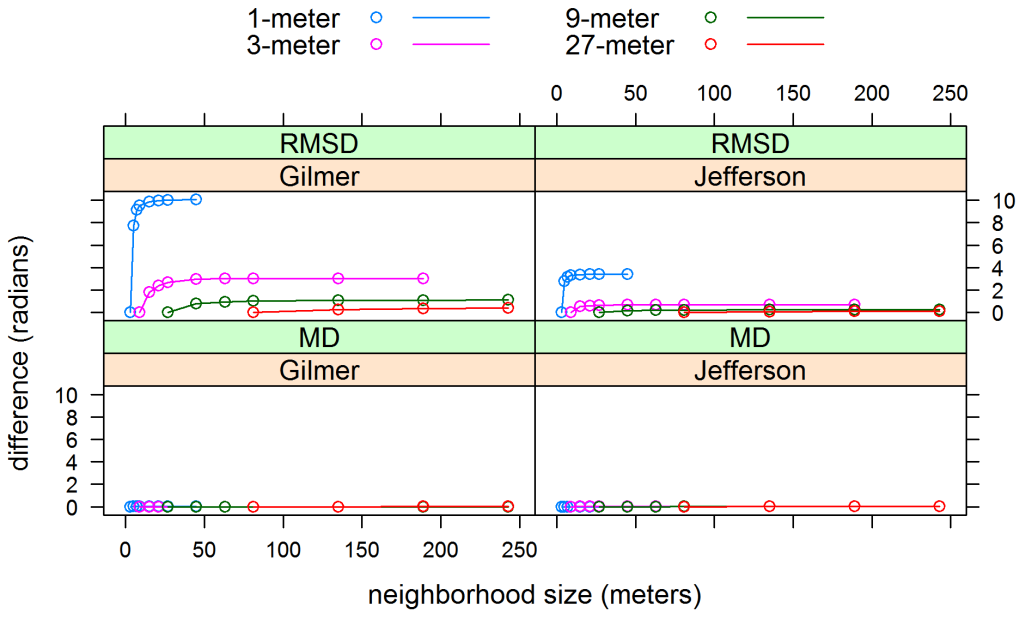

Figure 2.13: Plot of tangential curvature mean difference (MD) and root mean square difference (RMSD) goodness of fit measures for Gilmer and Jefferson Co. Slope gradient calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

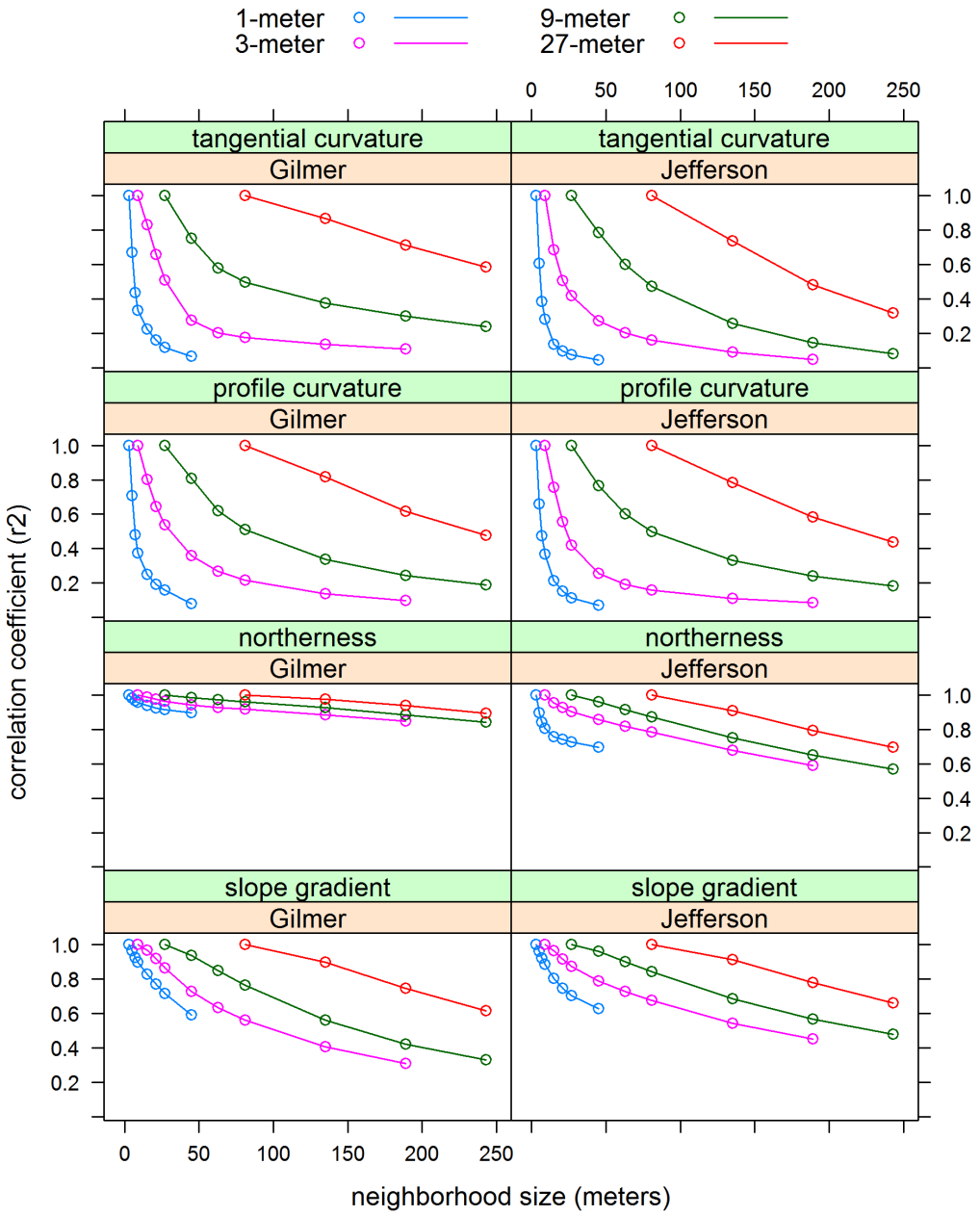

Figure 2.14: Plot of slope gradient, northerness, profile curvature, and tangential curvature correlation coefficient (r^2) goodness of fit measure for Gilmer and Jefferson Co. Slope gradient calculated using different grid (1, 3, 9, and 27-meters) and neighborhood sizes (3, 5, 7, 9, 15, 21, 27, 45, 63, 81, 135, 189, and 243-meters).

Case Study 2: Soil and LSP correlations response to grid and neighborhood size

The results show that the correlation between the soil properties and surface curvatures are the most sensitive to the effects of neighborhood size (Fig. 2.8-2.11). Their correlation with the soil attributes ranged from approximately 0 to 0.4, with an optimal neighborhood size range of 25-75 meters in most cases. That the correlations

between the soil properties and surface curvatures were optimized is important because slope curvatures have the strongest correlation coefficients in most cases.

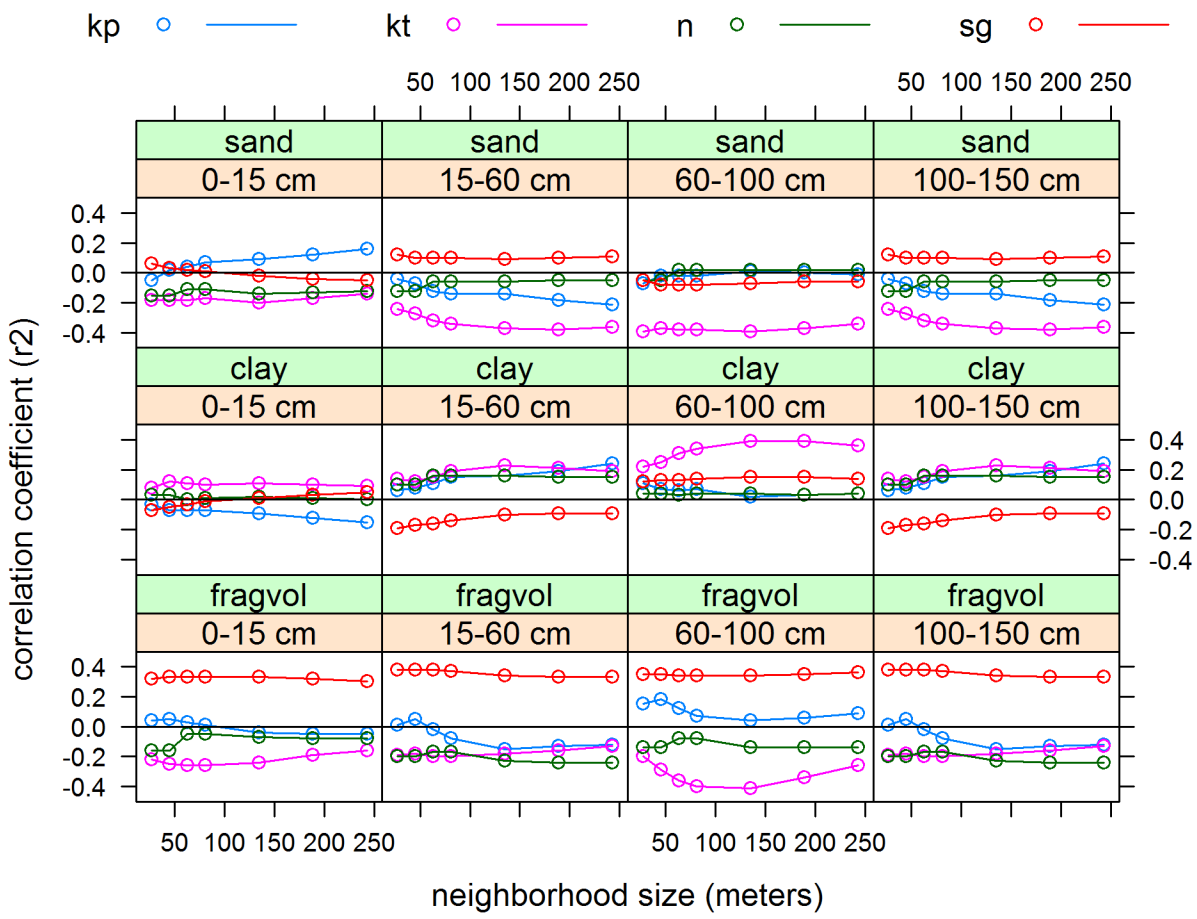

Figure 2.15: Case Study 2. Correlation coefficient vs. neighborhood size: rock fragments (fragvol), clay and sand. (kp = profile curvature, kt = tangential curvature, n = northerness, sg = slope gradient)

While significant correlations were also present between slope gradient and the soil properties, the effect of neighborhood size was negligible in most cases (Fig. 2.8-2.11). As for northerness, its correlation was minimal in almost all cases. The negligible effect of neighborhood size on the correlation between the soil properties and slope gradient may be explained by the results of the first case study. Increasing the neighborhood size used to derive LSP had less of an effect on slope gradient than on slope curvature (Figs. 2.4 and 2.5). Consequently, the correlation coefficients between soil properties and slope gradient do not appear to be sensitive to the effect of grid or neighborhood size.

In general, the effect of grid size showed similar trends and magnitudes as neighborhood size. However, whereas the effect of neighborhood size increased initially and then leveled off, the plots of grid size (Fig. 2.8 and 2.10) showed noticeable peaks and valleys. Thus the effect of neighborhood size seems to be more stable.

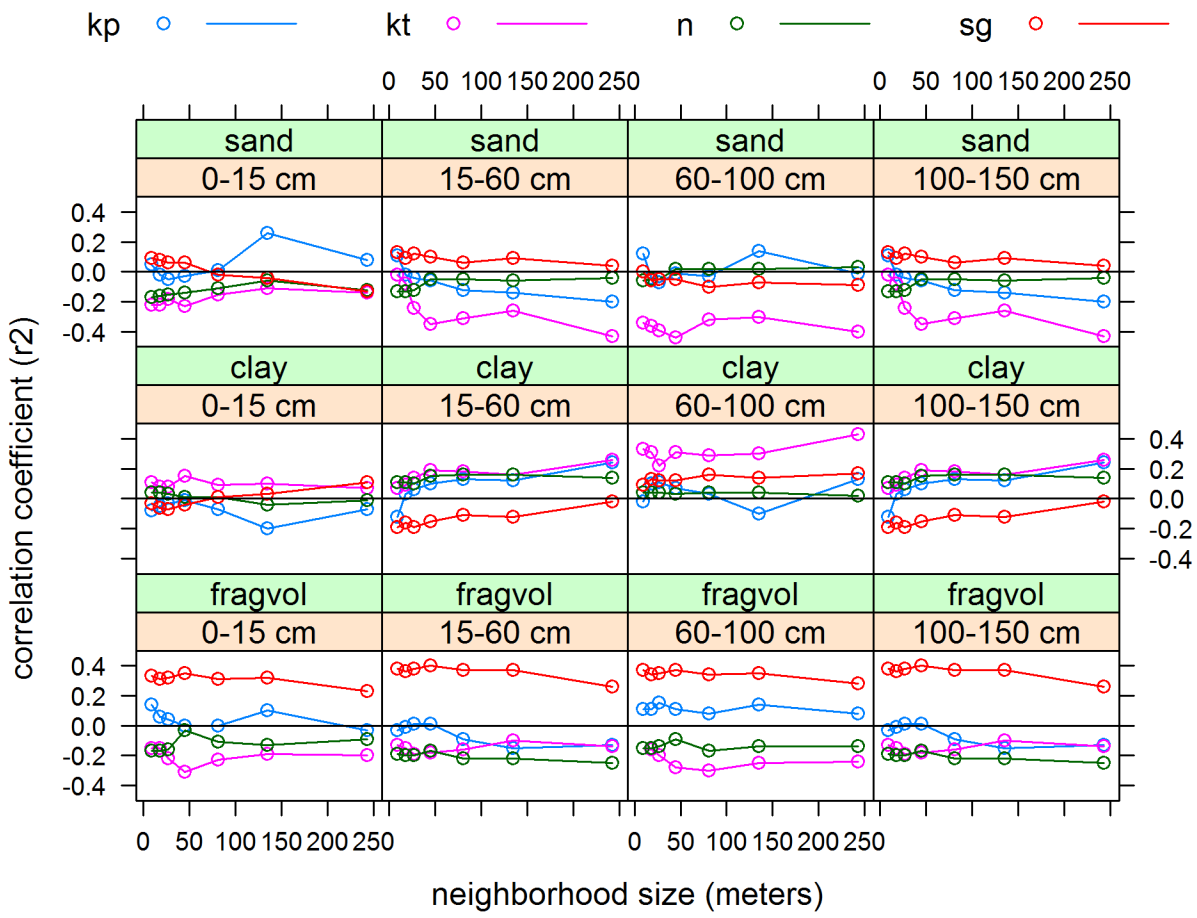

Figure 2.16: Case Study 2. Correlation coefficient vs. grid size: rock fragments (fragvol), clay and sand. (kp = profile curvature, kt = tangential curvature, n = northerness, sg = slope gradient).

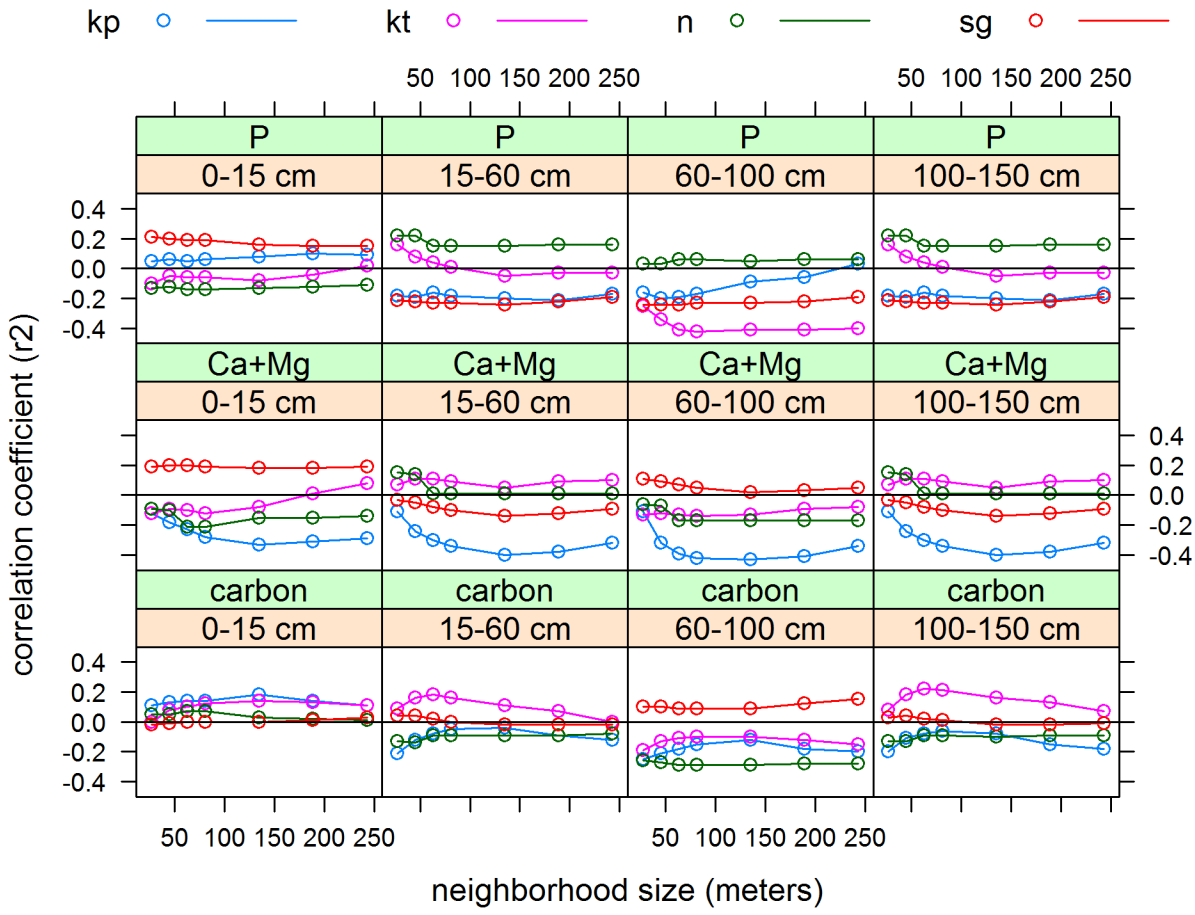

Figure 2.17: Case Study 2. Correlation coefficient vs. neighborhood size: carbon, calcium and magnesium (Ca+Mg), and phosphorus (P). (kp = profile curvature, kt = tangential curvature, n = northerness, sg = slope gradient)

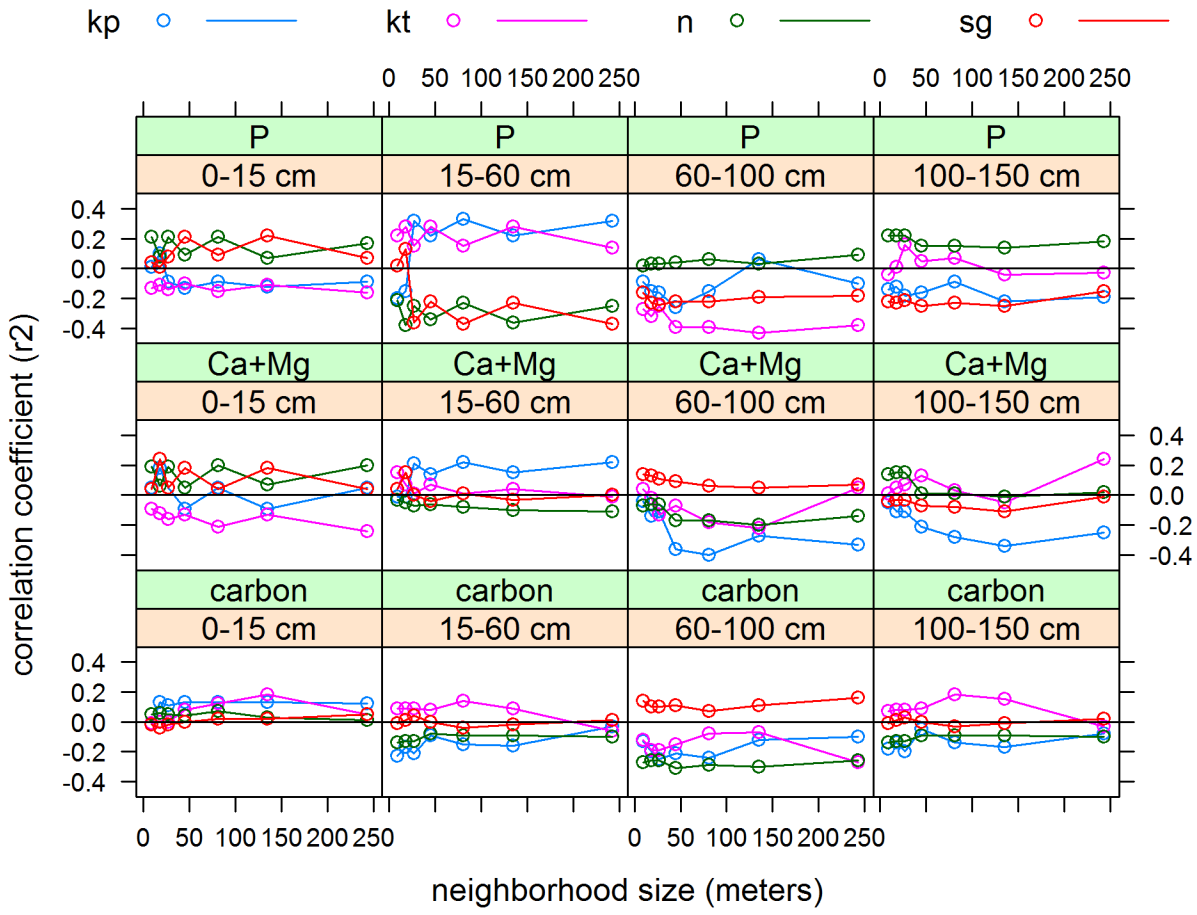

Figure 2.18: Case Study 2. Correlation coefficient vs. grid size: carbon, calcium and magnesium (Ca+Mg), and phosphorus (P). (kp = profile curvature, kt = tangential curvature, n = northerness, sg = slope gradient)

Conclusions

The spatial extent over which LSP are derived has a considerable effect on their representation of the land surface and correlation with soil attributes. In the first study described here, it was shown that using a larger neighborhood size has a similar effect as using a larger grid size, without the unnecessary loss of detail caused by using a larger grid size. Still the amount of detail provided by the smallest grid sizes was excessive and computationally demanding. To help determine what neighborhood size might be most appropriate for DSM, the simple exploratory procedures used here proved to be informative. Ultimately, the relative size of the landforms within the study area should serve as a guide. Within the second study described here, the correlation between the soil attributes and surface curvatures was optimized by varying the grid and neighborhood size. While no one spatial extent showed the strongest correlation in all cases, a common range of optimal neighborhood sizes occurred over a

range of 63-81-meters. Slope gradient also showed significant correlations with some of the soil properties, but was not sensitive to changes in neighborhood size. These results suggest that surface curvatures are the most sensitive to altering the neighborhood size used to calculate LSP and that curvature values poorly represent the land surface unless calculated over a neighborhood size commensurate with the size the landforms which they characterize.

Chapter 3
Statistical modeling of soil properties

Abstract

Digital soil mapping is a rapidly growing area of soil research that has great potential for enhancing soil survey activities and advancing knowledge of soil-landscape relationships. To date many successful studies have shown that geographic datasets can be used to model soil spatial variation. This study attempted to replicate that success in a West Virginia landscape, by predicting numerous physical and chemical soil properties for specific depth internals (0-15, 15-60, 60-100, and 100-150-centimeters), using generalized linear models (GLM) and geographic datasets. The area examined was the Upper Gauley Watershed on the Monongahela National Forest, which covers approximately 82,500 acres (33,400 hectares). Given this landscape diversity it was still possible to fit GLM which explained on average 38 percent of the adjusted deviance for rock fragment content, and exchangeable calcium and magnesium, and phosphorus. Some of the most commonly selected environmental predictors were slope curvatures, lithology types, and relative slope position indices. This seems to validate the prominence of these variables in theoretical soil-landscape models. Had the correlation between the soil properties and slope curvatures not been optimized by varying the spatial extent, it is likely that another less suitable LSP would have been selected.

Introduction

Soil is an integral component of terrestrial ecosystems, which makes ecosystem processes difficult to study in isolation of soil (Chapin et al., 2002). Therefore knowledge of the soil resource is informative to the explanation and management of many other natural resources. For this reason many industrialized nations have soil survey programs, which investigate and disseminate soil attribute and geographic information for a host of public uses. However, soil information is one of the most difficult environmental objects to observe and quantify, as it exists below the surface, varies in four dimensions (i.e. space and time), and is composed of three phases of matter (i.e solids, liquids, and gases). In addition to its sheer complexity, is the societal challenge presented to the soil sciences by the ever increasing pace of human environmental exploitation. This requires that the delivery of soil information be timely, relevant, coherent, and cost effective, so that it may be used to reduce the risks associated with environmental decisions (McKenzie et al., 2008). In order to keep pace with these societal challenges, the soil sciences have been spurred to integrate with other scientific disciplines and incorporate new concepts and technological developments.

Digital soil mapping (DSM) is a branch of pedometrics concerned with the prediction of spatial-temporal soil information by numerical and quantitative models (McBratney et al., 2003). Spatial prediction is typically achieved by one of two means, referred to as the spatial or *clorpt* approaches respectively, which in some circumstances

maybe mixed for optimal effect. The first approach interpolates predictions to new locations as a function of their distance from neighboring observations, by modeling the spatial dependence between observations with a variogram (i.e. geostatistics). This is an effective technique were the soil exhibits spatial correlation, is sampled at distances closer than the average range of spatial dependence, and can provide interpretable spatial statistics. However in most cases the soil varies so greatly over short distances, that this approach is generally considered impractical for mapping at small scales (i.e. large areas), which would otherwise require dense sampling. The second approach derives predictions of soil properties and/or types as a function of their relationship with environmental predictors, with knowledge-based (heuristic) or statistical models. This is an efficient technique when the environmental predictors are more easily attainable than the soil observations, and have a strong physical connection to the soil characteristics of interest (Gessler et al., 1995). The popularity the second approach stems from its strong theoretical framework, which is based the state factor model,

$$s = f(cl,o,r,p,t,...),$$

(Jenny 1941, 1980). Recently, McBratney et al. (2003) has expanded the original formulation of *clorpt* to include existing soil information (s) and spatial location (n), resulting in the acroynm, *scorpan* (note time, t, has been swapped with age, a). The incorporation of these additional factors recognizes their value for prediction. By utilizing both the spatial and *clorpt* approaches to DSM, researchers have modeled a variety of soil properties and types, over a range of scales, with varying degrees of accuracy and precision. At this point DSM has reached a point where it is considered ready for operational mapping (Burrough, 1993; MacMillan 2007; Hengl 2009), evidenced by many operational examples from around the world (Bui et al., 2003; MacMillan et al., 2005).

Operational DSM is now possible due to technological advancements, such as the introduction of digital geographic data-sets, numerical methods of analysis and increased computing capacity. However, DSM is advocated for both scientific and practical reasons. The scientific rationale for DSM is that its methods provide unbiased estimates of the central tendency of soil characteristics and their uncertainty, using primarily data driven methods. Also, such methods aid in the analysis and interpretation of complex datasets. In addition, DSM makes it possible produce continuous estimates of soil characteristics and their geographic distribution. Burrough (1993) provides a concise discussion of these issues. The practical rational for DSM concerns its potential to reduce the cost and time associated with soil mapping. To achieve these savings, DSM applies semi-automated methods to streamline the time intensive tasks of database manipulation, analysis, map production, as well as make efficient use of costly field data collection. Ideally the efficiencies gained by automation and improved efficiency could allow soil scientists additional time to improve the quality and quantity of the input data.

Hypothesis, motivation and objective

The hypothesis of this study was to that geographic datasets could serve as *scorpan* factors capable of predicting soil properties with statistical models. While this has been demonstrated in other landscapes, there was motivation to examine DSM applicability within a large and complex West Virginia landscape. Such studies are necessary to verify a given method's soundness, as different landscape represent different challenges. Also numerous studies show a fondness for modeling soil properties rather than soil taxonomic units, which is contrary to conventional soil mapping practice. Therefore the objective of this study was to develop statistical models of soil properties, using ancillary geographic datasets as predictors.

Methods

Study area

The area examined for this study was the Upper Gauley Watershed, which totals approximately 82,700 acres, and occurs within the Monongahela National Forest of West Virginia (Fig 3.1). This location occurs within the Eastern Allegheny Plateau and Mountains, which is dominated by very steep side slopes, with narrow ridge tops and valleys (USDA-NRCS, 2006). Elevation within the watershed ranges from 658 to 1,435 meters. Precipitation is evenly distributed throughout the year, occuring as both rain and snow, with seasonal averages ranging from 130 to 165 centimeters (51 to 65 inches) (PRISM Climate Group, 2010). This has resulted in soils with udic and perudic soil moisture regimes (Delp, 1998; Flegel, 1998). Temperatures are cold during the winter and warm during the summer, with seasonal averages ranging from 6.5 to 10 degrees Celsius (44 to 50 degrees Fahrenheit) (PRISM Climate Group, 2010). This has resulted in mesic and frigid soil temperature regimes (Delp, 1998; Flegel, 1998). Frigid soils are typically found at elevations generally greater than 852 meters, while perudic soils are found within hollows. Vegetation is dominated by three forest types, including deciduous hardwoods at lower elevations, coniferous at higher elevations, and small patches of mixed types where deciduous and coniferous forest meet. The geology of the watershed is composed of sedimentary rocks that include Pennsylvanian aged sandstones (Pottsville Group) and Mississippian aged shale mixed with fine grained sandstone (Mauch Chunk Group). The Pottsville Group overlies the Mauch Chunk Group, which has been exposed in the floodplains and valleys of the eastern portion of the watershed.

The general distribution of the soils within the watershed as summarized by STATSGO2 (U.S. General Soil Map) soil map units which include: the Gilpin-Laidig (s8817), Trussel-Simoda-Mandy-Gauley (s8852), and Shouns-Cateache-Belmont (s8823) associations (Figure 3.1). Both the Gilpin-Laidig and Trussel-Simoda-Mandy-Gauley associations are formed on the Pottsville sandstone within the western and central portions of the study area respectively, while the Shouns-Cateache-Belmont association is formed on the Mauch Chunk shale on the eastern side of the watershed. The Gilpin-Laidig and Trussel-Simoda-Mandy-Gauley association differ based on elevation

and slope gradient, with Trussel-Simoda-Mandy-Gauley association present at elevations generally greater than 852 meters and on steeper slopes.

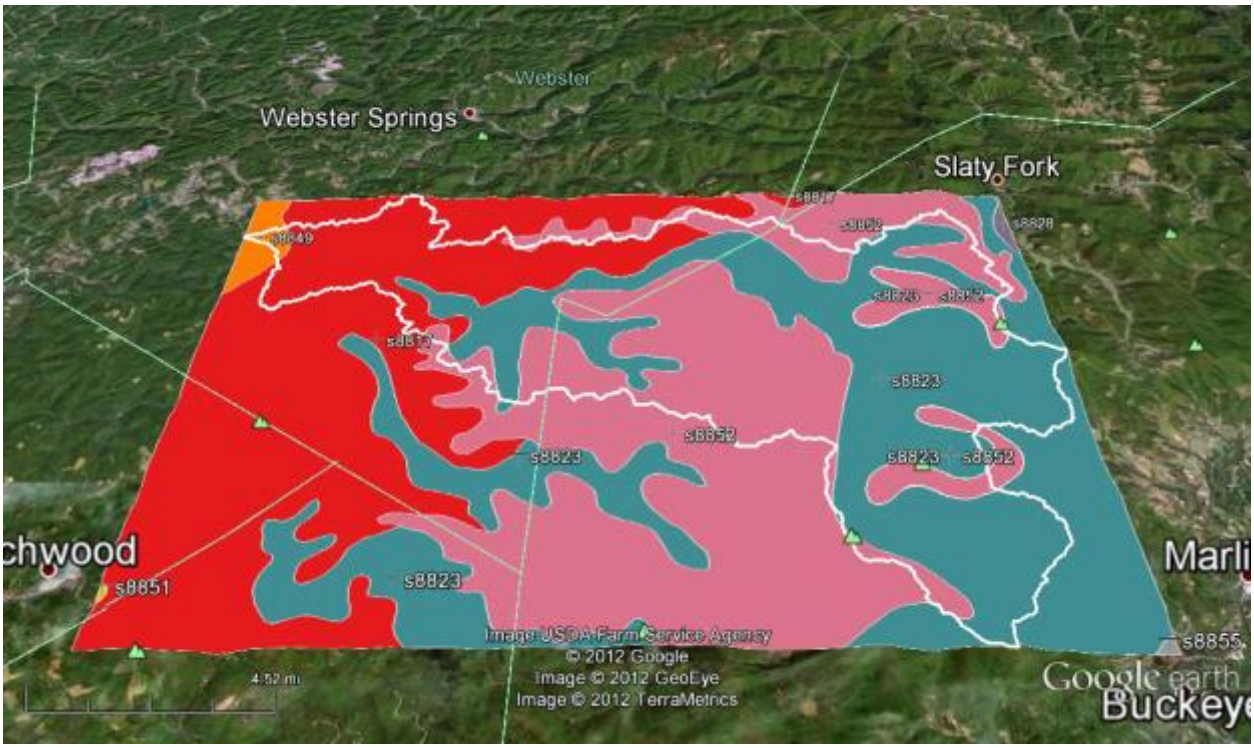

Figure 3.1: U.S. General Soils Map (STASTGO2) of the Upper Gauley Watershed. (Gilpin-Laidig (s8817), Trussel-Simoda-Mandy-Gauley (s8852), and Shouns-Cateache-Belmont (s8823) soil associations).

Table 3.1: Summary of named soil series from the U.S. General Soils Map (STASTGO2) of the Upper Gauley Watershed. (Gilpin-Laidig (s8817), Trussel-Simoda-Mandy-Gauley (s8852), and Shouns-Cateache-Belmont (s8823) soil associations).

STASTGO2 Symbol	Soil Series	Soil Classification	Depth class	Drainage class
s8817	Gilpin	Fine-loamy, mixed, active, mesic Typic Hapludults	moderately deep	well
	Laidig	Fine-loamy, siliceous, active, mesic Typic Fragiudults	very deep	well
s8852	Trussel	Fine-loamy, mixed, semiactive, frigid Aeric Fragiaquepts	very deep	poorly
	Simoda	Fine-loamy, mixed, semiactive, frigid Typic Fragiudepts	deep & very deep	moderately well
	Mandy	Loamy-skeletal, mixed, active, frigid Typic Dystrudepts	moderately deep	well
	Gauley	Loamy-skeletal, siliceous, superactive, frigid Typic Haplorthods	moderately deep	well
s8823	Shouns	Fine-loamy, mixed, semiactive, mesic Typic Hapludults	very deep	well
	Cateache	Fine-loamy, mixed, active, mesic Ultic Hapludalfs	moderately deep	well
	Belmont	Fine-loamy, mixed, active, mesic Typic Hapludalfs	deep	well

Soil Sampling and Analysis

Site allocations

Ninety seven points were sampled. To allocate the sampling locations a design-based stratified-random sampling strategy was used, similar to that of McKenzie and Ryan (1999). The stratifying variables used were geology, elevation, and stream power index (SPI). Geology was stratified into the watershed's three lithology types (sandstone, shale/sandstone, and shale). Elevation was separated into three quantile classes, and SPI (Tarboton, 2004) was separated into 5 quantile classes. The intersection of these variables created 45 unique strata within which to sample. In order to avoid sampling extraneous features, exclusion rules were used to avoid roads (buffered to 20 meters), streams (buffered to 10 meters), developed areas, and patches of strata smaller than 4,000 square meters (approximately one acre). In addition private land and the Cranberry Wilderness Area were excluded from sampling.

To randomly allocate the sampling sites, five patches from each stratum were randomly selected using Hawth's Analysis tools for ArcGIS (Beyer, 2004). Within each patch, a random site (x,y) was generated. This produced 225 potential sites. In order to decide which sites to visit, they were randomly ranked between 0 and 1. The first two sites with the lowest rank were selected for each stratum, and successive sites where selected if field observation found that they violated the exclusion rules.

Soil profile and site descriptions

Upon field sampling, each site was located using a Trimble GeoXM GPS unit. At each site, two 32-foot (9.75-meters) transects centered on the random coordinates where established. One transect was oriented directly downslope, and the other across slope and to the left. The transect length was chosen to represent the resolution of a 10-meter grid. These transects served to estimate the percentage of surface rock fragments coverage (sfragcov). The location of the soil pit was placed within the area intersected by the transects, at a location that seemed representative, but as close to the intersection of the transects as possible.

At each site, a soil pit was hand dug to a depth of at least 140-cm or bedrock, and described according to standard procedures (Schoeneberger et al., 2002). The percent volume of rock fragments (fragvol) was visual estimated by comparison with standardized percent surface area charts. From each soil horizon, two 300-gram grab samples were taken, one for laboratory analysis, and another for characterization and archiving by the National Forest Soil Scientist. The percent of surface rock fragments at each site was determined by thrusting a spade through the O horizon along 1 foot intervals of both transects, and recording the number of instances a rock fragment was struck. This procedure is similar to that used by Jenkins (2002), although he multiplied the total percentage of surface rock fragments by an arbitrary factor of 0.6, to account for the amount of void space he assumed to occur between the rock fragments.

Laboratory analysis

The grab samples which were collected from each horizon, were analyzed for percent carbon, pH, extractable metals (Aluminum (Al), Phosphorus (P), Calcium (Ca), Magnesium (Mg), and Manganese (Mn)), and particle size. To prepare the samples for analysis they were first air-dried, ground, and passed through a 2-mm sieve. Soil carbon (C) was analyzed twice on approximately 0.095-g of soil using a Leco Truspec CHN elemental analyzer. Soil pH was measured on approximately 10-mL of soil in a 20-mL 0.01 *M* $CaCl_2$ solution (Lierop, 1990). The extractable Al, P, Ca, Mg, and Mn were measured on approximately 5 cm^3 of soil using a Mehlich 1 extracting solution (0.025 *N* H2SO4 + 0.05 *N* HCl)(Mehlich, 1953), and analyzed on a PerkinElmer ICP (inductively coupled plasma) optical emission spectrometer (Optima 2100 DV). The extractable metals and C were converted to a mass per unit area (kg/ha), by multiplying their concentrations by their horizon thickness, bulk density, and a correction factor for the amount of rock fragments. Because preliminary analysis showed Ca and Mg to be highly correlated (i.e. 0.89), only their sum was analyzed (i.e. Ca + Mg). Clay was measured on approximately 40-g of soil using the hydrometer method, while sand was estimated by wet sieving (Gee and Bauder, 1986). Because bulk density was not measured for any of the horizons within this study, it was estimated from 15 soil profiles sampled by Sponaugle (2005) in the adjacent Cranberry watershed. From these soil profiles a regression equation was developed that explained over 73 % of variance using the % carbon and middle depth of each soil horizon as predictive variables.

$$\text{Bulk density} = 1.303 + -0.077(\text{carbon}) + 0.007(\text{depth})$$

To summarize the soil attributes for analysis, the soil horizons from each soil profile were aggregated into four depth intervals (0-15, 15-60, 60-100, and 100-150 cm) by taking a weighted average using the aqp R package (Beaudette et al., 2012). This is a common approach for modeling soil properties vertical anisotropy (McKenzie and Austin, 1993; McKenzie and Ryan, 1999; Park and Vlek, 2002; Malone et. al, 2009, Odgers et. al, 2012). The specific depth intervals used in this study were chosen because they seemed to correspond with inflections observed in depth plots of the soil properties (Figure 3.2). For the purposes of this study, the thickness of the O horizons was not included, because no laboratory analysis was performed.

Table 3.2: Summary of the soil properties.

Soil property & depth interval	Definition (units)	Measurement
sfragcov fragvol 0-15-cm fragvol 15-60-cm fragvol 60-100-cm fragvol 100-150-cm	Surface cover of rock fragments (%) Volume of rock fragments (%)	64-ft transect Visual estimation (Schoeneberger et al., 2002)
clay 0-15-cm clay 15-60-cm clay 60-100-cm clay 100-150-cm	weight of clay for < 2-mm particle size fraction (%)	Hydrometer (Gee and Bauer, 1986)
sand 0-15-cm sand 15-60-cm sand 60-100-cm sand 100-150-cm	weight of sand for < 2-mm particle size fraction (%)	Sieve (Gee and Bauer, 1986)
C 0-15-cm C 15-60-cm C 60-100-cm C 100-150-cm	soil organic carbon (kg/ha)	Leco Truspec CHN elemental analyzer
pH 0-15-cm pH 15-60-cm pH 60-100-cm pH 100-150-cm	activity of H ions	0.01 *M* $CaCl_2$ solution (Lierop, 1990)
Ca+Mg 0-15-cm Ca+Mg 15-60-cm Ca+Mg 100-cm Ca+Mg 100-150-cm	sum of exchangeable Calcium and Magnesium (kg/ha)	Mehlich 1 extracting solution (Mehlich, 1953)
P 0-15-cm P 15-60-cm P 60-100-cm P 100-150-cm	exchangeable Phosphorus (kg/ha)	Mehlich 1 extracting solution (Mehlich, 1953)
Al 0-15-cm Al 15-60-cm Al 60-100-cm Al 100-150-cm	exchangeable Aluminum (kg/ha)	Mehlich 1 extracting solution (Mehlich, 1953)

Table 3.3: Summary of environmental predictors.

State factors (abreviation)	**Definition/Significance** (units)	**Algorithm/Reference**
Climate		
Mean annual air temperature (tmean)	temperature (degrres Fahrenheit)	PRISM Climate Group, 2010
Total annual precipitation (ppt)	precipitation (inches)	PRISM Climate Group, 2010
Parent material (geo)		
Lithology type: sandstone (ss), shale (sh), or sandstone & shale (sssh)	parent material (scale 1:250,000)	West Virginia Geological and Economic Survey, 1968
Land surface parameters - Geometric		
Elevation (z)	height above mean sea level (feet)	
Slope gradient (sg) with a 7x7 window size	down slope rate of change (percent)	Wood, 1996; GRASS, 2012
Profile curvature (kp) with a 7x7 window size	down slope curvature (radians)	Wood, 1996; GRASS, 2012
Tangential curvature (radians) (kt) with a 7x7 window size	across slope curvature (radians)	Wood, 1996; GRASS, 2012
Multiresolution valley bottom flatness index (mrvbf) > 0.1 equals reclassified as valleys	index of flatness and lowness	Gallant and Dowling, 2003
Land surface parameters - Regional		
Specific catchment area (sca)	meters^2/grid size (meters)	Seibert and McGlynn, 2007
Topographic wetness index (twi)	soil saturation index, ln(sca/sg)	Moore et al., 1991
Stream power index (spi)	soil erosion index, ln(sca*sg)	Moore et al., 1991
Catchment height (ch)	average upslope height (meters)	Moore et al., 1991
Mid-slope position (zms)	relative topographic position (percent)	Bohner and Antonic, 2009
Normalized height (zhn)	height above valleys (percent)	Bohner and Antonic, 2009
Annual solar radiation or insolation (sr_a)	amount of incoming solar energy (kWh/mA2)	Wilson and Gallant, 2000
Remotely sensed imagery		
Tassel cap component 1 (spring (tc1_s), leafon (tc1_lo), leafoff (tc1_lf))	soil brightness	USGS, 2006
Tassel cap component 2 (spring (tc2_s), leafon (tc2_lo), leafoff (tc2_lf))	greeness	USGS, 2006
Tassel cap component 3 (spring (tc3_s), leafon (tc_lo), leafoff (tc3_lf))	wetness (moisture)	USGS, 2006

* For the purpose of spatial analysis all other layers were resampled to 15-meters.

Environmental predictors

A database of environmental predictors (Table 3.1) was developed using the GIS, SAGA (SAGA, 2012). All of the geospatial data were freely available online, but required some preprocessing. The DEM came from the Statewide Addressing and Mapping Board (SAMB) and USGS Elevation Conversion Project. This DEM was produced using triangular irregular network (TIN) interpolation of mass points and break lines sampled from stereo pairs of aerial photography. The original resolution of the DEM was 3-meters, but it was resampled to 15-meters using mean aggregation, in order to minimize the abundance of triangular artifacts inherited from the interpolation. The results of Chapter 2 showed this to be one of the most highly correlated grid sizes (Figure 2.16 and Figure 2.18). The remotely sensed imagery used was selected from the MRLC Landsat 7 TM+ Scene Library, which contains images from three seasons (i.e. spring, leafon, and leafoff) that have been terrain corrected and converted to at-sensor reflectance (USGS, 2006). From this library, the tasseled cap components were selected as predictors due to their well-established physical interpretation and data reduction properties.

Statistical analysis

Exploratory data analysis

Prior to predictive modeling, exploratory analysis of the soil properties was performed to assess their distributions and interrelationships. The distribution of the soil properties was assessed by examining normal quantile-quantile (Q-Q) and plots of their mean and quartiles (i.e. 0.25 and 0.75 percentiles) with depth. Skewness was visually assessed by how much the distribution of a given soil property at a specific depth interval diverged from a theoretical normal distribution. To correct for skewness, transformations of the soil properties were made where necessary using natural logarithm (i.e. log) and square root (i.e. sqrt) functions.

To assess the interrelationships between the soil properties and observations, biplots and scatterplots were visually examined for each depth interval. Only the 1st and 2nd principal component (PC) were interpreted, as they usually explain the majority of variance. To assess the importance of each of the components a table of their standard deviations and variance proportions was examined.

Biplots are useful graphical technique as they display the multivariate (e.g. overall) structure amongst the variables and observations in one plot. The variables themselves are displayed as arrows. The direction of an arrow quantifies the variables loading (i.e. one minus the correlation) to the PC, and length quantifies the proportion of variance explained for a PC. While the correlation between variables is not specifically addressed within the biplots, if the correlation between variables is strong, the angle between variable's arrows can imply correlation (i.e. small angles approximate strong correlations). The distance between the observations quantifies their similarity

(i.e. approximate Mahalanobis distances) to each other, while their location relative to the arrows relates to the abundance of each variable.

Predictive modeling

The statistical models evaluated for prediction were generalized linear models (GLM), regression trees (RT), and random forest (RF) using the glm (R Core Team, 2013), rpart (Thereneau et al., 2013), and randomForest (Liaw and Wiener, 2002) R packages. The GLM were fit using forward variable selection. Due to strong co-linearity between the numerous predictors, automated step-wise selection was avoided. Instead variables were sequentially selected by evaluating their p-values, Akaike's Information Criteria (AIC), adjusted D^2, cross-validated (CV) root mean square error (RMSE), and whether the relationships made sense. The GLM were also assessed for constant variance, normality, linearity, co-linearity, outliers and influential points. In order to utilize soil profiles that were shallower than a given depth internal (e.g. 60-90-cm, rather than 60-100-cm), the thickness of each observation was used as prior weights. The stopping criteria for the RT, was the minimum number of splits that fell within one standard deviation of the CV residual sum of squares (RSS). The RF was fit using the default regression parameters of 500 trees and a terminal node size of 5. Also individual trees within the RF were grown using a bootstrap sample of two thirds of the observations and one third of the predictors. To validate the statistical models 10-fold CV was used.

Results and Discussion

Exploratory data analysis

Examination of the Q-Q plots showed Ca+Mg, P, Al, and percent rock fragment volume (fragvol) to be skewed for all depth intervals, while C was only skewed below 60-cm. To normalize the skewed variables, a natural logarithm transformation (log) was sufficient, except for Al which required a square root transformation (sqrt). While the Q-Q plots are not display below, the skewed nature of these soil properties is also evident by examining the median and quartiles displayed in Table 2.1 and Figure 3.2.

The depth plots in comparison showed the distribution of each soil properties median, and 25^{th} and 75^{th} quartiles for 1-cm depth increments. Overall depth trends showed near constant clay, sand and C content with depth, and an increase for fragvol, pH, Ca+Mg, P, and Al with depth. Particularly striking is the low values and narrow range of pH, which are considered extremely acid. This is to be expected considering the area's acid rain, and the acidic properties of the Pottsville sandstone. Curiously there is a small decrease in clay content with depth. This trend is the reverse of what would be expected, considering that numerous soil series within the study area are characterized as having an argillic horizon, which should display a clay increase or bulge below the topsoil. The fact

that the extractable metals increase drastically with depth, is probably do to leaching, which is enhanced by acid rain, and nutrient extraction by plants.

Table 3.4: Statistical summary of the soil properties for each depth interval.

Soil property & depth interval	Summary					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
sfragcov	0	17	41	41	59	100
fragvol 0-15-cm	0	5	10	20	28	90
fragvol 15-60-cm	0	8	15	22	28	90
fragvol 60-100-cm	0	10	23	28	40	92
fragvol 100-150-cm	0	20	37	37	51	95
clay 0-15-cm	4	17	23	24	31	43
clay 15-60-cm	4	16	21	20	25	34
clay 60-100-cm	2	15	18	18	21	37
clay 100-150-cm	5	14	17	19	23	43
sand 0-15-cm	7	25	35	37	47	91
sand 15-60-cm	9	31	36	40	48	92
sand 60-100-cm	8	35	42	42	51	76
sand 100-150-cm	3	34	42	42	52	75
C 0-15-cm	0	3	4	4	6	11
C 15-60-cm	1	2	3	3	4	8
C 60-100-cm	1	2	2	3	3	9
C 100-150-cm	1	2	3	3	4	12
Ca+Mg 0-15-cm	3	63	112	373	228	4081
Ca+Mg 15-60-cm	28	85	131	448	293	6028
Ca+Mg 100-cm	21	118	230	705	543	7988
Ca+Mg 100-150-cm	25	158	369	1378	1054	11080
P 0-15-cm	0	1	2	3	4	13
P 15-60-cm	0	1	2	4	4	54
P 60-100-cm	0	1	2	5	5	45
P 100-150-cm	0	1	3	9	8	76
Al 0-15-cm	31	747	1257	1323	1749	4790
Al 15-60-cm	31	1374	2050	2337	3162	5613
Al 60-100-cm	359	1655	2355	2720	3412	7127
Al 100-150-cm	264	1393	2417	3106	4229	10940

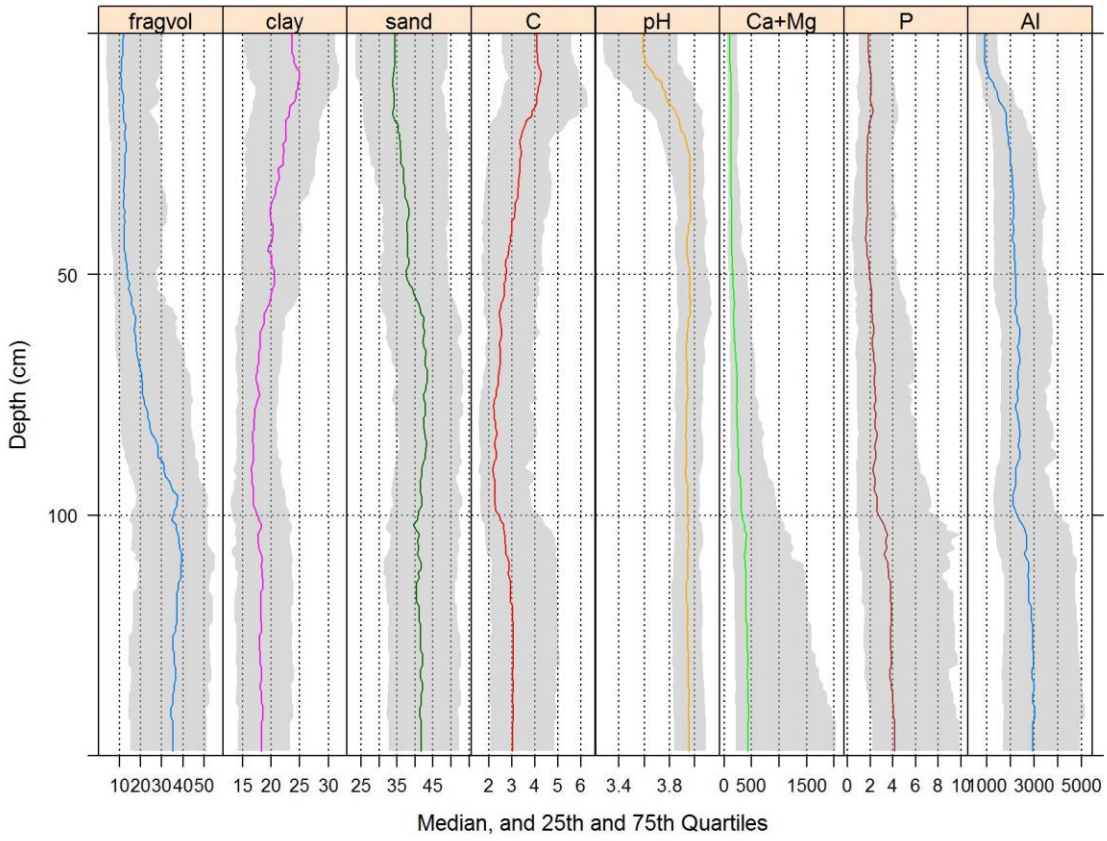

Figure 3.2: Depth plot of the mean and lower (0.25th) and upper quartiles (0.75th) for rock fragments (fragvol)(% volume), clay (% weight), sand (% weight), carbon(C)(kg/ha), pH (unitless), Calcium and Magnesium (Ca+Mg) (kg/ha), Phosphorus (P)(kg/ha), and Aluminum (Al)(kg/ha).

A summary of the PC analysis that provides the PC scores and loadings for the biplots is presented in Table 3.2, which contains the standard deviation and proportion of variance explained by each PC. It shows that each depth interval has a similar proportion of variance explained by each PC, each of whose cumulative proportion increases gradually which suggests little data redundancy in the dataset. For each depth interval the 4th PC is the first to fall below one standard deviation. Also, for each depth interval the first two components contain 50-60 percent of the cumulative variance. Ninety percent of the cumulative variance isn't achieved until the 5th PC.

The biplots are shown in Figure 3.3, and the scatterplots are shown in Figures 3.4 and 3.5. They display the interrelationships between the soil properties and observations for each depth interval. While by their nature biplots are complex figures which attempt to condense large amounts of information, the relations between the different depth intervals creates an additional layer of complexity. In general the biplots show that the first PC is dominated by the textural properties (i.e. fragvol, clay and sand) of the soil for each interval. Interestingly none of the textural properties are appreciably associated with Ca+Mg; rather Ca+Mg is mostly correlated with C and P for all depth

intervals, while it's relation to Al appears to be nonlinear according to the smoothing splines (Figure 3.4 and 3.5). This is somewhat surprising considering the importance of clay, whose extensive surface area is generally attributed to soil's nutrient capacity. Again in this instance the extreme acidity of the soil is likely responsible. Instead the clay is likely saturated with Hydrogen (H), Al and P, which may explain clay's moderate to low correlation with Al ,pH and P. Predictably clay and sand are negatively correlated for most properties, while sand and fragvol are correlated. The C surfaces by contrast are also likely saturated Al, given their moderately positive correlations.

The overall distribution of soil observations within the biplots reveals their similarity to each other and the soil properties. Notable is a lack of clustering amongst the observations. Rather the observations from a diffuse cloud, with a central concentration of observations. Thus the PC of each depth interval is defined by a relatively few observations that deviate from the central concentration of observations. No trend is apparent in the distribution of diffuse soil observations between depth intervals. While not specifically addressed in this study, it suggests that for the soil properties examined here, the dissimilarity between the soil taxonomic units within this area is not great.

Table 3.5: Importance of the principal components for each depth interval.

0-15 cm								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.67	1.36	1.03	0.90	0.80	0.66	0.49	0.39
Proportion of Variance	0.35	0.23	0.13	0.10	0.08	0.05	0.03	0.02
Cumulative Proportion	0.35	0.58	0.71	0.82	0.90	0.95	0.98	1.00
15-60 cm								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.61	1.25	1.20	0.96	0.80	0.68	0.48	0.39
Proportion of Variance	0.33	0.19	0.18	0.12	0.08	0.06	0.03	0.02
Cumulative Proportion	0.33	0.52	0.70	0.82	0.89	0.95	0.98	1.00
60-100 cm								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.57	1.30	1.23	0.95	0.73	0.68	0.48	0.44
Proportion of Variance	0.31	0.21	0.19	0.11	0.07	0.06	0.03	0.02
Cumulative Proportion	0.31	0.52	0.71	0.82	0.89	0.95	0.98	1.00
100-150 cm								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.64	1.38	1.21	0.86	0.77	0.53	0.47	0.33
Proportion of Variance	0.34	0.24	0.18	0.09	0.07	0.04	0.03	0.01
Cumulative Proportion	0.34	0.58	0.76	0.85	0.92	0.96	0.99	1.00

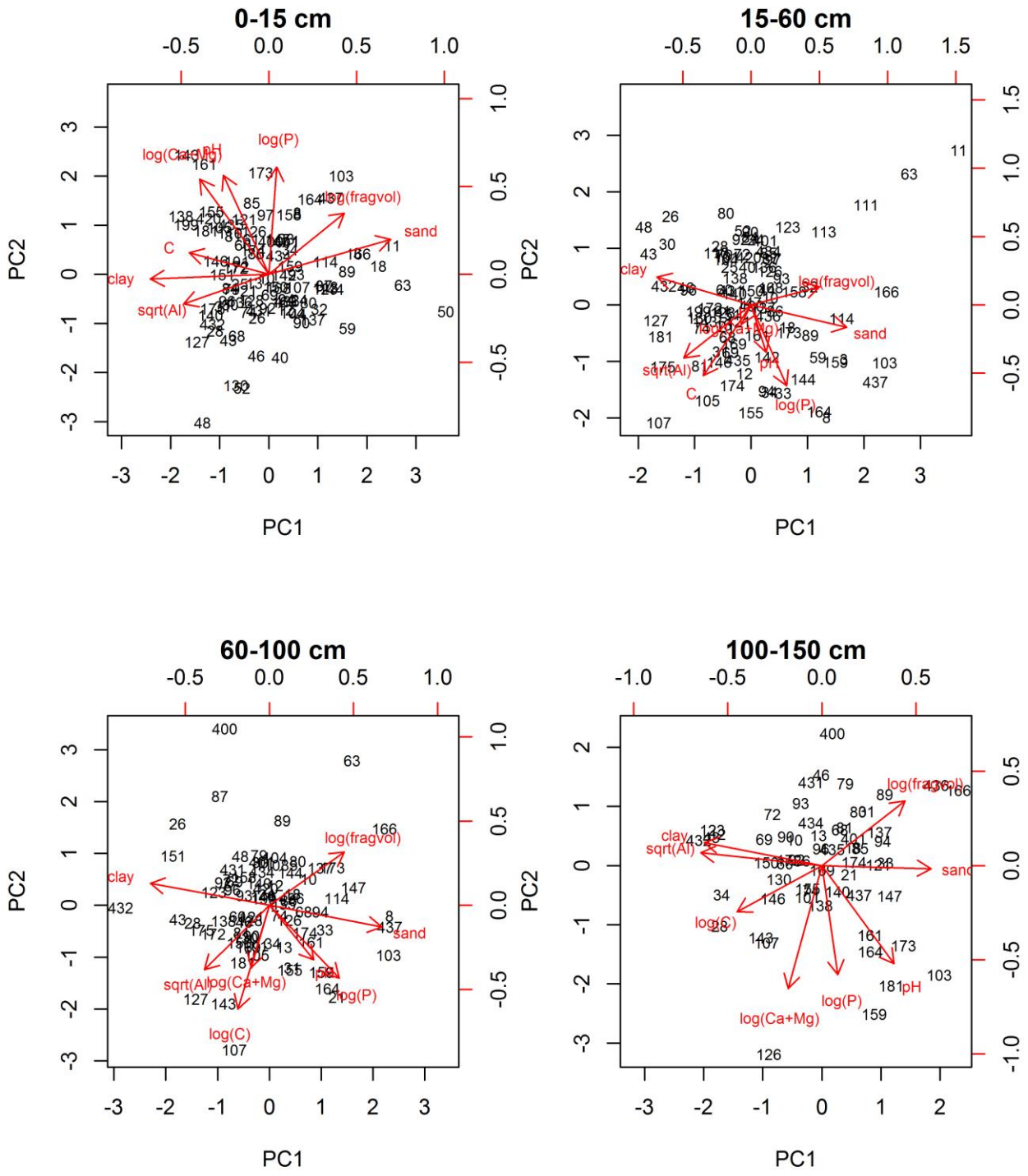

Figure 3.3: Biplots of the soil properties for each depth interval. The bottom and left axes represents the standardized component scores (i.e. soil observations), while the top and right axes represent one minus the standardized component loadings (i.e. soil properties).

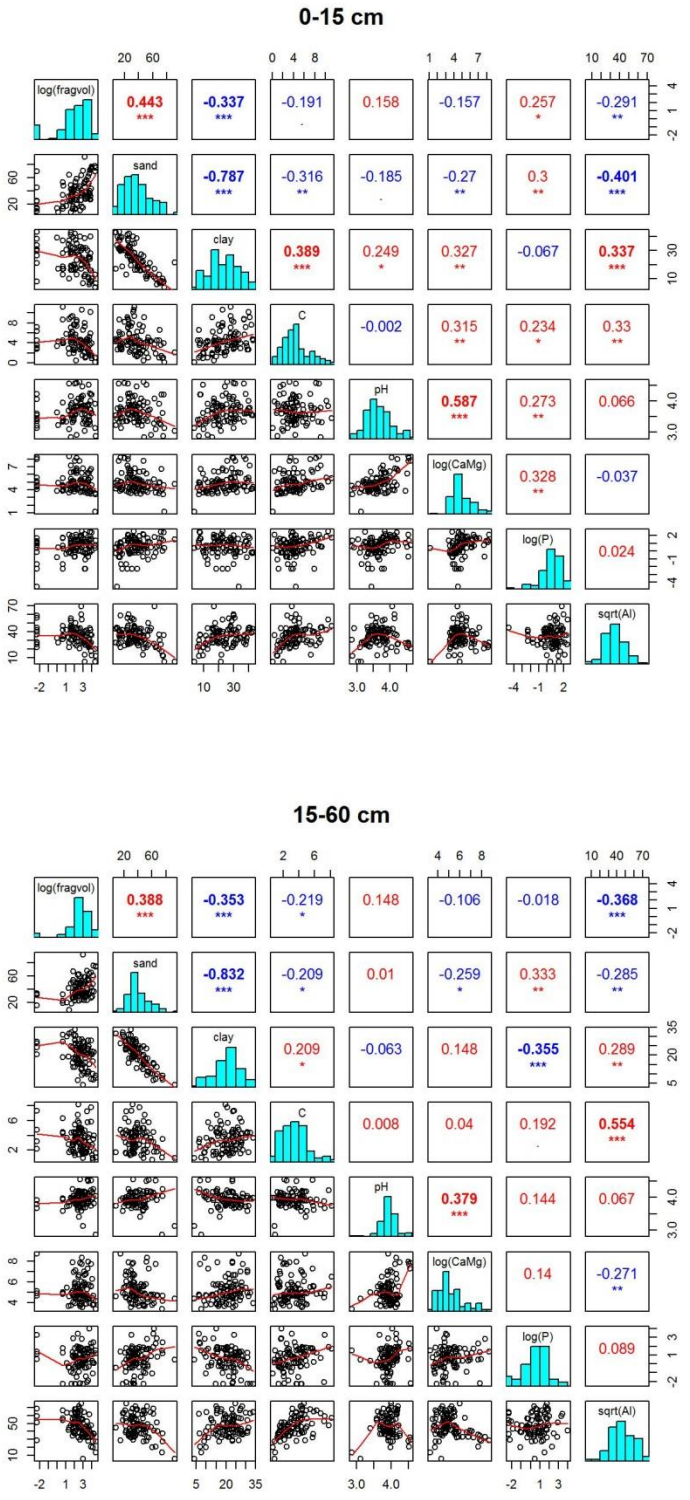

Figure 3.4: Scatterplots, histograms, correlation matrix of depth intervals 0-15 cm and 15-60 cm. Smoothing line fitted to the scatterplot. Significance levels: 0.1 (.), 0.05 (*), 0.01 (**), 0.001 (***). Units: fragvol log(% volume), clay (% weight), sand (% weight), C (kg/ha), pH (unitless), Ca+Mg log(kg/ha), P log(kg/ha), and Al log(kg/ha).

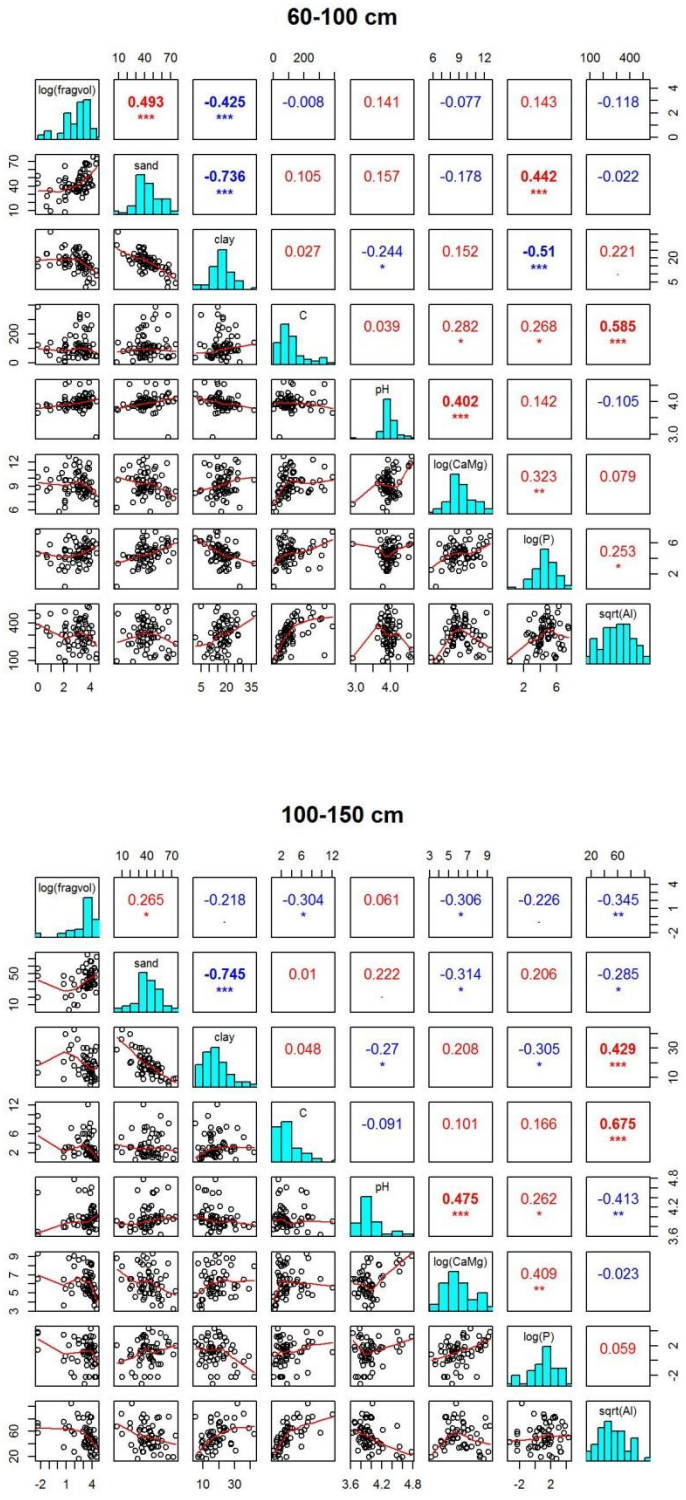

Figure 3.5: Scatterplots, histograms, correlation matrix of depth intervals 60-100 cm and 100-150 cm. Smoothing line fitted to the scatterplot. Significance levels: 0.1 (.), 0.05 (*), 0.01 (**), 0.001 (***). Units: fragvol log(% volume), clay (% weight), sand (% weight), C log(kg/ha), pH (unitless), Ca+Mg log(kg/ha), P log(kg/ha), and Al log(kg/ha).

Predictive modeling

A separate generalized linear model (GLM) was constructed for each soil property and depth interval. The results of which are summarized in Table 3.6 and Table 3.7. Overall the GLM outperformed the regression trees (RT) and random forests (RF), therefore only GLM were constructed. Discussions of the various aspects of the GLM are presented under separate headings. Due to the number of models created, no attempt was made to interpret each individual model individually. Instead the interpretation of the results focused mainly on the relation between models and environmental predictors included.

Examination of the RT showed that in some cases the cross-validated (CV) error increased after only one split. The RF were similarly unable to identify splits in some cases when validated with the bootstrap samples withheld when growing the trees, termed the 'out of bag' (OOB) data. Regression trees and RF are common data mining techniques that have become popular within DSM, in part because their automated fitting procedures, particularly when multiple predictors are present. It maybe that the RT and RF underperformed because they generally require sizeable datasets and make inefficient use of continuous predictors. In such circumstances Maindonald and Braun (2007) suggest that parametric models, such as GLM, may be a better alternative as they make assumptions of the form of the data. Thus these results are significant as they suggest it may be necessary in some cases for multiple models to be evaluated, particularly less automated methods which require more statistical assumptions, and statistical knowledge from the analyst.

Accuracy

The accuracy of the GLM varied between the soil properties and depth intervals examined (Table 3.4). In comparison the models of fragvol, Ca+Mg, and P were the most accurate, with their adjusted deviance squared (adj D^2) ranging from 0.12 to 0.61, and averaging 0.38. Also the cross-validated RMSE was similar to the resubstituted RMSE, which demonstrates the stability of GLM fit. Amongst these models, the accuracy was generally highest for the 15-60 and 60-100-cm depth intervals. The models of the other soil properties performed less well, but identified as similar selection of interpretable predictors. The generally low accuracy of the GLM maybe the result of a number of factors, such as unspecified interactions or insufficient predictors. Alternatively it maybe that only a narrow range in a particular soil property was observed; such as C and clay. For example, the RMSE of clay was 5-8 percent, which only slightly worse than laboratory methods such as the hydrometer (Gee and Bauder, 1986). However, Ryan et al. (2000) suggests that spatial models most likely are only capable of explaining 70 percent of the variation in soil properties, with models explaining less than 50 percent being most common. Therefore it is more likely that the soil properties of this landscape maybe dominated by random fluctuations. In which case some measure of central tendency and dispersion maybe sufficient to characterize their variability. By comparison, this is the approach practiced in soil survey. For example, the soil map units in the watershed are generally composed of

consociations, which are dominated by a single soil series, and characterized by a high, low, and representative value for each soil horizon.

GLM variance functions

In addition to their predictive utility, the GLM also provide addition information on the error distribution of the soil properties. For example, while the preliminary exploratory analysis suggested that many of the soil properties response required transformations in order to achieve a normal distribution, the GLM diagnostic plots showed that the error terms (or variance functions) of many also required additional transformations in order to satisfy the linear models assumption that the errors are normally distributed. While the variance functions of the physical soil properties were mostly normal (i.e. constant), the chemical soil properties variance increased with the mean (mu). This required the application of either a Gamma or quasipoisson GLM. The consequence of this is that the variability of those soil properties increases in conjunction with their estimated mean. Interestingly, the error distribution of fragvol and C changed with depth.

Within DSM the application of GLM with non-normal error terms is unusual, except for binomial GLM (i.e. logistic regression) which are used for classification instead of regression (McKenzie and Ryan, 1993; Gessler et al., 1995; Beaudette and O'Geen, 2009b). Generally most researchers make use of Gaussian GLM with transformations of the response (Thompson et al., 1997; McKenzie and Ryan, 1999; Park and Vlek, 2002). Park and Vlek (2002) for example compared several statistical models including GLM, and various soil properties for five depth intervals. They only reported using transformations of the response variables and interaction terms. Regardless Lane (2002) has demonstrated the application of GLM with alternative variance functions in soil science in order to satisfy the assumption of linear models. This may be due as Lane (2002) suggests to most soil scientists unfamiliarity or misunderstanding of GLM.

Table 3.6: Summary of the GLM constructed for each soil property and depth interval.

Soil property & depth interval	Formula	Family	Link function	Variance function	Outliers	df	Adjusted D^2	RMSE	cv RMSE
sfragcov	spi + tc2_lf + sr_a	gaussian	identity	constant		93	0.36	21.51	22.04
fragvol 15-cm	spi + tc2_lo + geo + zms	quasipoisson	log	mu		91	0.28	18.16	20.31
fragvol 60-cm	spi + tc2_lo + geo	quasipoisson	log	mu	142	89	0.33	16.37	18.05
fragvol 100-cm	spi + tc2_lo + geo	gaussian	log	constant	137,400	72	0.43	15.2	16.99
fragvol 150-cm	spi + tc2_lo + sg	gaussian	log	constant		56	0.24	21.38	23.88
clay 15-cm	zms + tc2_lf + kt	gaussian	identity	constant	155,199	89	0.27	8.11	8.43
clay 60-cm	kt + zms	gaussian	identity	constant	114,436	89	0.11	6.16	6.33
clay 100-cm	kt + zms	gaussian	identity	constant		75	0.16	5.34	5.57
clay 150-cm	zhn + tmean + tc2_lo	gaussian	identity	constant	123	55	0.29	6.14	6.57
sand 15-cm	kt + zms	gaussian	identity	constant	11,114	90	0.16	16.05	16.96
sand 60-cm	kt + zms	gaussian	identity	constant	114	74	0.25	12.01	12.72
sand 100-cm	kt	gaussian	identity	constant	114	75	0.24	12.22	12.62
sand 150-cm	zhn + tc3_s	gaussian	identity	constant	123,33	54	0.32	11.06	11.6
C 15-cm	tc2_s	gaussian	identity	constant		93	0.04	2.33	2.36
C 60-cm	tc3_lo	gaussian	identity	constant	12,96,105 ,107,155	87	0.16	1.22	1.25
C 100-cm	zhs+sh	quasipoisson	log	mu		73	0.2	1.64	1.76
C 150-cm	1	quasipoisson	log	mu		58	0.15	2.18	2.24
Ca+Mg 15-cm	geo + kp + tc2_s + zms	Gamma	log	mu^2	181	88	0.43	598.47	649.8
Ca+Mg 60-cm	geo + kp + tc2_s	Gamma	log	mu^2	181	88	0.47	482.07	499.38
Ca+Mg 100-cm	geo + kp + zms	Gamma	log	mu^2		73	0.44	773.37	941.64
Ca+Mg 150-cm	geo + kp	Gamma	log	mu^2		55	0.27	1991.33	2234.01
P 15-cm	zms + tmean + kt	quasipoisson	log	mu		85	0.12	2.38	2.47
P 60-cm	mrvbf + zms + kt	quasipoisson	log	mu	3	83	0.61	2.31	2.63
P 100-cm	mrvbf + geo + zms + kt + kp	quasipoisson	log	mu	68	63	0.58	3.3	3.62
P 150-cm	sg + geo	quasipoisson	log	mu	126,140,1 61	48	0.42	5.78	6.76
Al 15-cm	geo + kt	quasipoisson	sqrt	mu	33	90	0.08	687.84	722.13
Al 60-cm	kp + tmean	quasipoisson	sqrt	mu		91	0.12	1172.1	1220.27
Al 100-cm	kp + tmean + kt	quasipoisson	sqrt	mu		74	0.18	1455.34	1507.87
Al 150-cm	kp + tmean + geo	quasipoisson	sqrt	mu		54	0.13	2017.85	2263.27

Outliers

While fitting the GLM, various observations stuck out as outliers. These observations were highlighted in residual plots, and removed from the GLM prior to their final fitting. For most models this required removing only a single observation, in some cases it was necessary to remove several. No apparent cause for the outliers was determined.

Environmental predictors

A comparison of the environmental predictors used in each GLM is presented in Table 3.5. While each GLM contained a unique combination of environmental predictors, similar variable combinations occurred amongst the soil properties and depth intervals examined (Table 3.5). Also, of particular interest is that the sign and magnitude

of the GLM coefficients for the same soil property at different depth intervals were similar. This suggests that in many cases the landscape processes that shape the spatial distribution of specific soil properties operate similarly over a range of depths. However differences were noted. Generally the effect of LSP appears to diminish below 100-cm or be absent. This result agrees with those of Park and Burt (2002), who also found the effect of LSP to decrease with depth. For example, with sfragcov and fragvol where SPI was included for all depths intervals, the slope of SPI decreased with depth but was similar between 0 and 100-cm. By comparison, the effect of geology increased with depth for Ca+Mg.

Amongst the GLM, the most common environmental predictors selected were the slope curvatures, lithology types (geo), and relative slope positions. This seems to validate the prominence of these variables in theoretical soil-landscape models. Surprisingly, the hydrologic LSP (e.g. stream power index and topographic wetness index) (Table 3.2) did not appear commonly in the GLM, considering their popularity in DSM. Commonly hydrologic LSP are selected because they condense multiple LSP (e.g. landscape processes) into one index, negating the need to incorporate two predictors. However, while the hydrologic LSP commonly ranked high during the variable selection process, their effect was negated when the slope curvatures and mid-slope position were introduced, due to the high correlation between the slope curvatures and the hydrologic LSP. Had the correlation between the soil properties and slope curvatures not been improved in Chapter 2 it is likely that the variable selection process would have preferentially selected hydrologic LSP instead. Thus again this demonstrates the importance of neighborhood size when calculating slope curvatures.

New to this study was the incorporation of the LSP mid-slope position position (zms), normalized landscape position (zhn), and multi-resolution valley bottom index (mrvbf). These LSP are all recent additions to the DSM toolbox. Typically within DSM landscape position is estimated with hydrologic LSP, as they distinguish uplands from lowlands (e.g. hydrologic connectivity or topological relations). This is important because while land surfaces may have similar shapes, their landscape position will determine their level of exposure to additions of sediment and water from upslope areas. Thus, given the inclusion of these new landscape position predictors into the GLM, it appears they contain some new useful information other than that already provided by existing hydrologic LSP; perhaps particularly in a steep landscape such this study area. Within the GLM, zms was the most commonly included. Low values of zms distinguished mid-slope positions (e.g. backslopes) from their surrounding summits and valleys. For example, the zms coefficients (Table 3.5) from the clay and sand GLM demonstrate that mid-slope positions have more clay and less sand relative to summits and valleys. This is likely the result of enhanced clay formation on slopes from throughflow. By comparison, zhn provides an estimate of landscape position above valleys, with 0 percent characterizing valleys and 100 percent characterizing summits. This position measure is particularly similar to twi and spi, but provides a more general spatial distribution as it involves an iterative slope-based modification of catchment area. For example, the zhn coefficients from the clay and sand GLM demonstrate that between 100-150-cm clay increases with increasing height above valleys and correspondingly decrease in sand, albeit

at different rates. This could be interpreted to show that in stable landscape positions such as summits, clay is translocated to lower depths, while on mid-slopes positions greater clay contents are found higher in the soil profile. Lastly, the mrvbf is designed to distinguish valleys from uplands by using a nonlinear threshold transformation of slope gradient and elevation percentile over a range of scales. While only a minority of the study area is characterized by valley floors, the mrvbf proved to be a useful predictor were soil properties within the floodplain deviated from the overall trend. It may be in floodplains that separate GLM slope coefficients exist for the other predictors, but there were only sufficient observations within the floodplain to capture the effect of mrvbf.

Besides geology and the LSP, the multi-temporal tasseled cap components also were included in numerous GLM. The most commonly selected tassel caps components were the 2nd (tc2). The tassel cap components are similar to principal components, but represent a fixed linear transformation has been found useful for agricultural monitoring. The 2nd tasseled cap is associated with the greenness of vegetation (Table 2.1), and is strongly correlated with the normalized vegetation index (NDVI) which indicates vegetation abundance. Interestingly, no one season (i.e. spring, leafon, or leafoff) of tc2 appeared to be the best predictor of all the soil properties. These different seasonal images are poorly correlated therefore it is difficult to speculate on a causal relationship with the soil properties. In general the spring (tc2_s) and leafon (tc2_lo) tc2 images appear to correspond with a greater density of coniferous trees, which typically occur at higher elevations, while the leafoff (tc2_lf) tc2 images appear to correspond with south aspects and areas without trees. At a minimum these results suggest the importance of examining images from multiple seasons for relationships with soil properties. The use of satellite imagery is not uncommon in DSM, but in reviewing the literature this author was only able to find a few recent examples of this approach (Sylvain et al., 2012). Sylvian et al. (2012) provides a noteworthy example were a normalized band ratio of wet versus dry seasons was used in an agricultural setting from Quebec, Canada.

Given that some environmental predictors were included in several GLM, they would make good potential stratifying variables for future work in this landscape setting, such as the slope curvatures, geology, mid-slope position, and tc2 components. Although in the future it might be more efficient to distribute the sites using a multivariate Latin hypercube design (Minasny and McBratney, 2006) using R (Roudier, 2011). In this study the stratifying variables were treated as blocking effects in an ANOVA design, however for predictive purposes this distinction was ignored. There was an attempt to force these variables into the GLM with the combination of other predictors (Table 3.5) in order to see if they contributed to the deviance, but only geo was significant. By themselves some of the stratifying variables levels were significant for some models, but did not contribute substantially to the deviance to be useful for predictive purposes.

Table 3.7: Summary of the GLM coefficents for each soil property and depth interval.

Soil property & depth interval	Intercept	Geology		Land surface parameters								Tasseled cap components					PRISM
		sh	sssh	kp	kt	sg	mrvbf	spi	zms	zhn	sr_a	tf2_lf	tc2_lo	tc2_s	tc3_s	tc3_lo	tmean
sfragcov	-138.15							8.85			-0.02	1.41					
fragvol 15-cm	4.53	-0.57	-0.11					0.33	0.71				-0.02				
fragvol 60-cm	4.98	-0.44	-0.10					0.34					-0.02				
fragvol 100-cm	5.49	-0.44	-0.08					0.34					-0.03				
fragvol 150-cm	4.40					0.01		0.19					-0.02				
clay 15-cm	84.73				3.49				-14.43			-0.53					
clay 60-cm	23.03				3.57				-5.26								
clay 100-cm	19.55				3.88				-4.40								
clay 150-cm	125.56									13.20			-0.17				-1.80
sand 15-cm	30.87				-11.33		18.31		11.67								
sand 60-cm	39.46				-12.77				8.34								
sand 100-cm	43.60				-13.25												
sand 150-cm	74.06									-28.34					-0.23		
C 15-cm	12.37													-0.09			
C 60-cm	16.50															-0.12	
C 100-cm	-0.65	-0.25	0.23							-0.43				0.02			
C 150-cm	1.24	-0.31	0.35														
Ca+Mg 15-cm	11.02	1.16	0.73	-1.63					-0.98					-0.06			
Ca+Mg 60-cm	9.01	0.94	0.74	-2.20										-0.04			
Ca+Mg 100-cm	5.89	1.35	0.47	-2.05					-0.97								
Ca+Mg 150-cm	5.62	2.04	1.22	-1.54													
P 15-cm	-6.57				-0.21				0.88								0.15
P 60-cm	0.46				-0.65		1.88		1.07								
P 100-cm	0.44	0.82	0.29	-0.68	-0.48		2.47		0.74								
P 150-cm	2.54	0.88	0.94			-0.02											
Al 15-cm	32.42	6.19	2.62		3.63												
Al 60-cm	146.35			13.47													-2.12
Al 100-cm	210.34			15.32	5.72												-3.43
Al 150-cm	269.44	-9.04	-0.40	22.40													-4.54

Spatial predictions

An example of the spatial prediction and standard error (SE) for each depth interval of Ca+Mg are presented in Figures 3.6 and 3.7. Within these figures the increase in Ca+Mg with depth is apparent, as previously indicated in the depth plot of Ca+Mg (Figure 3.7). Also apparent is an increase in the standard error (SE) with depth. By examining the Figures, the influence of the predictors can be seen as expressed by the GLM coefficients (Table 3.4). For example the overlying strata of Pottsville sandstone has visibly less Ca+Mg, as would be expected considering its acidic composition. Also readably visible for all depth intervals is the effect of profile curvature, which indicates that convex positions are lower in Ca+Mg, relative to concave positions.

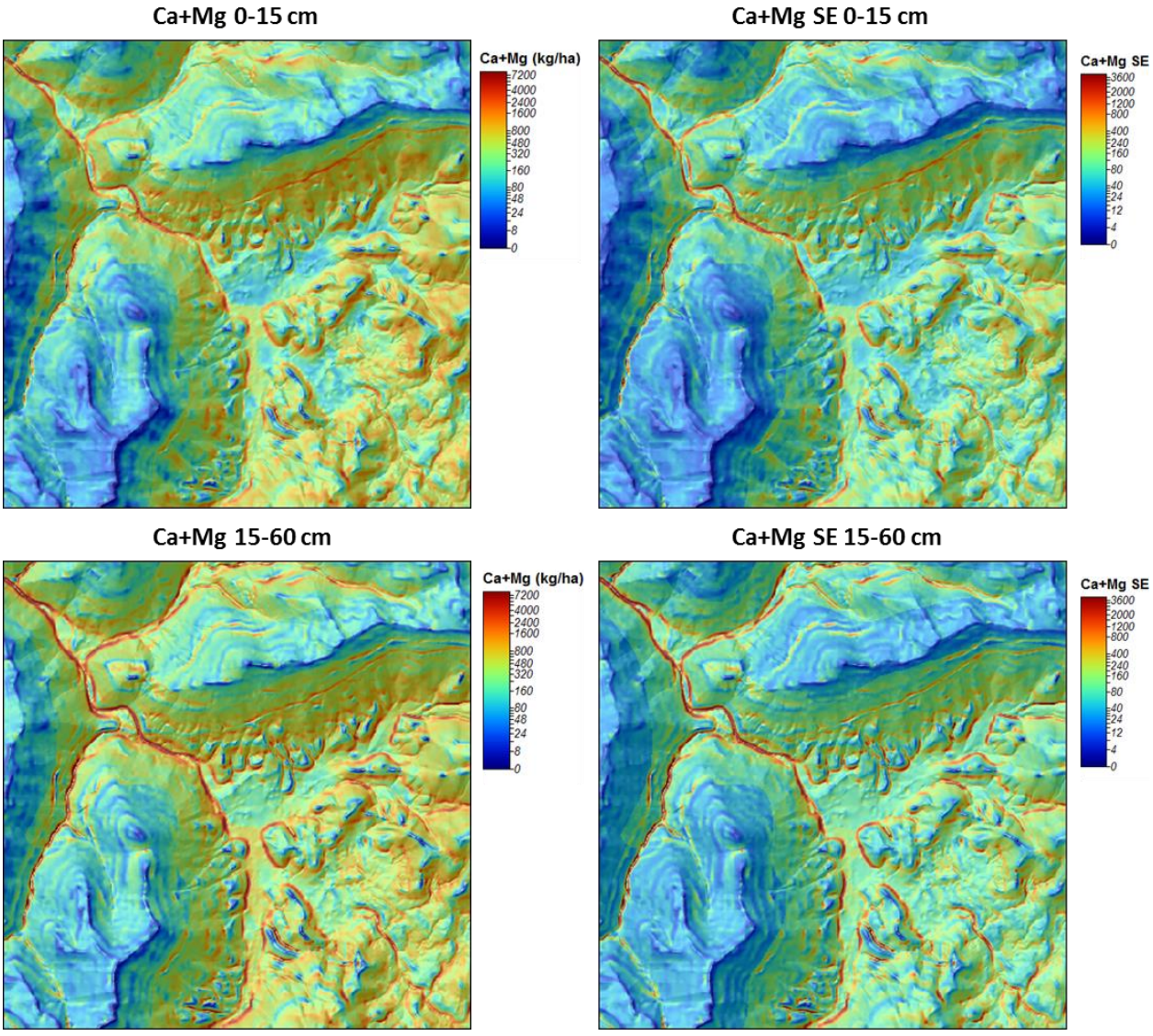

Figure 3.6: Spatial prediction of Ca+Mg (mg/ha) and standard error (SE) for the 0-15 and 15-60-cm depth intervals.

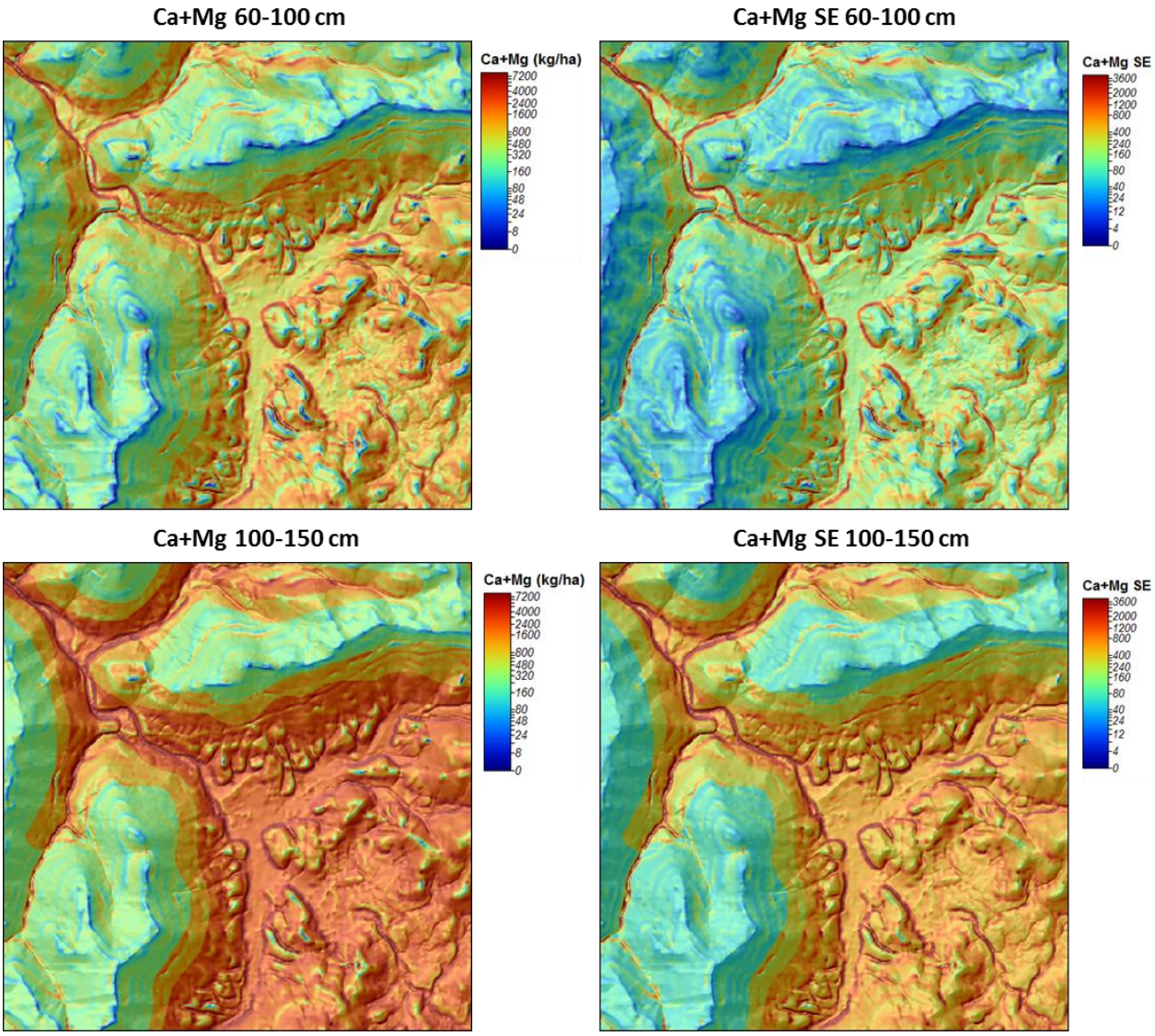

Figure 3.7: Spatial prediction of Ca+Mg (mg/ha) and standard error (SE) for the 60-100 and 100-150-cm depth intervals.

Conclusions

Numerous spatial models of soil properties for several depth intervals were constructed using generalized linear models (GLM) and environmental predictors within a West Virginia landscape. The GLM demonstrated a range of accuracies (i.e. 0 to 61 % adjusted deviance explained) for various soil properties and depth intervals. The most successfully modeled soil properties were rock fragment content, and exchangeable calcium and magnesium, and phosphorus (averaging 38 % adjusted deviance explained). Considering the narrow quartile range of values observed for clay, sand, and carbon their models might be adequate. Otherwise they might best be modeled by a simple depth function for the entire watershed. Exploratory data analysis and the GLM diagnostics indicated that several of the soil properties had non-normal response distributions, while in addition all of the chemical soil prop-

erties also had non-normal error term distributions. While transformations of the response variable is common in digital soil mapping, there appears to be little precedence for using non-normal error terms (i.e. Gamma and quasipoisson family GLM), besides the binomial family (i.e. logistic regression). Amongst the GLM, the most common *scorpan* factors selected were the slope curvatures, lithology types, and relative slope positions. This seems to validate the prominence of these variables in theoretical soil-landscape models. Had the correlation between the soil properties and slope curvatures not been improved in Chapter 2 it is likely that another less suitable hydrologic land surface parameter (e.g. stream power index), would have been selected instead. Within digital soil mapping there is a preference to produce models of soil properties, rather than taxonomic units or other predetermined classes. This study demonstrates that this is possible, but the question remains whether it is practical for soil survey agencies, considering the difficulty in fitting numerous separate models for each soil property and depth interval. If this is the case, then the alternative question remains as to how to incorporate the results of digital soil property maps into soil survey maps. Particularly those soil properties that are not typically captured in soil survey databases. It may simplify matters to use statistical models that have automated fitting procedures, such as regression trees and random forest, but this study generally found GLM to be more accurate and interpretable. There is great value to soil taxonomic units as they categorize soil information for the sake of comprehension. However the process of un-categorizing soil classes back into soil properties is likely to introduce error. Therefore future work could examine the tradeoff between the efficiency of taxonomic units, versus directly estimating soil properties.

References

Arnold, R.W., 2006. Soil Survey and Soil Classification. In: Grunwald, S. (Ed.), Environmental Soil-Landscape Modeling – Geographic Information Technologies and Pedometrics. CRC, Boca Raton, FL, pp.38-57.

Beaudette, D.E. and A.T. O'Geen, 2009a. Soil-Web: An online soil survey for California, Arizona, and Nevada. Computers & Geosciences. 35:2119:2128.

Beaudette, D.E. and A.T. O'Geen, 2009b. Quantifying the aspect effect: an application of soil radiation modeling for soil survey. Soil Sci. Soc. Am. J. 73(4)1345:1352.

Beaudette, D.E., P. Roudier, and A.T.O'Geen, 2012. Algorithms for quantitative pedology: a toolkit for soil scientists. Computers & Geosciences. 52:258-268.

Bell, J.C., R.L. Cunningham, and M.W. Havens, 1992. Calibration and validation of a soil-landscape model for predicting soil drainage class. Soil Sci. Soc. Am. J. 56:1860-1866.

Bell, J.C., R.L. Cunningham, and M.W. Havens, 1994. Soil drainage probability mapping using a soil-landscape model. Soil Sci. Soc. Am J. 58:464-470.

Beyer, H. L., 2004. Hawth's Analysis Tools for ArcGIS. Available at http://www.spatialecology.com/htools.

Bishop, T.F.A., and A.B. McBratney, 2001. A comparison of prediction methods for creation of field-extent soil property maps. Geoderma 103:149-160.

Bishop, T.F.A., and B. Minasny, 2006. Digital Soil-Terrain Modeling: The Predictive Potential of Uncertainty. In: Grunwald, S. (Ed.), Environmental Soil-Landscape Modeling – Geographic Information Technologies and Pedometrics. CRC, Boca Raton, FL, pp.185-208.

Brus, D.J., and J.J. de Gruijter, 1997. Random sampling or geostatistical modeling? Choosing between design-based and model-based sampling strategies for soil. Geoderma 80:1-59 (with discussion).

Bui, E.N., and C.J. Moran, 2003. A strategy to fill gap in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia. Geoderma 111:21-44.

Buol, S.W., F.D. Hole, R.J. McCracken, and R.J. Southard, 1997. Soil Genesis and Classificatioin 4th edition. Iowa State University Press, Ames, Iowa.

Burrough, P.A., 1993. Soil variability: a late 20th Century view. Soil Fertilizers 56:529-562.

Breiman, L., 1996. Bagging predictors. Machine Learning 26:123-140.

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

Claessens, L., G.B.M. Heuvelink, J.M. Schoorl, A. Veldkamp, 2005. DEM resolution effects on shallow landslide hazard and soil redistribution modelling. Earth Surface Processes and Landforms 30:461-477.

Chapin, F. S. III, P.A. Matson, and P.M Vitousek, 2011. Principles of terrestrial ecosystem ecology. Springer, New York, NY, pp.546.

Chaplot, V., C. Walter, and P. Curmi, 2000. Improving the soil hydromorphy prediction according to DEM resolution and available pedological data. Geoderma 97:405-422.

Chaplot, V., F. Darbouz, H. Bourennane, S. Leguedois, N. Silvera, and K. Phachomphon, 2006, Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density. Geomorphology 177:126-141.

Daily, G.C., P.A. Matson, and P.M. Vitousek, 1997. Ecosystem services supplied by soil. In: Daily, G. (Ed.), Natures Services: Societal Dependence on Natural Ecosystems. Island Press, Washington, DC, pp.113-132.

Delp, C.H., 1998. Soil Survey of Webster County, West Virginia, U.S. Department of Agriculture, Natural Resource Conservation District. US. Gov. Print, Washington, DC.

Evans, I.S., 1972. General geomorphometry, derivations of altitude and descriptive statistics. In: Chorley, R.J. (Ed.), Spatial Analysis in Geomorphology. Harper & Row, New York, NY, pp.17-90.

Flegel, D.G., 1998. Soil Survey of Pocahontas County, West Virginia, U.S. Department of Agriculture, Natural Resource Conservation District. US. Gov. Print, Washington, DC.

Florinsky, I.V., 1998. Accuracy of local topographic variables derived from digital elevation models. Int. J. GIS 12(1):47-61.

Florinsky, I.V. and G.A. Kuryankova, 2000. Determination of grid size for digital terrain modelling in landscape investigations – exemplified by soil moisture distribution at a micro-scale. Int. J. GIS 14(8):815-832.

Florinsky, I.V., R.G. Eilers, G.R. Manning, and L.G. Fuller, 2002. Prediction of soil properties by digital terrain modeling. Environmental Modelling & Software 17:295-311.

Gallant, J.C., and T.I. Dowling, 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39(12):1347-1361.

Gee, G.W. and J.W. Bauder, 1986. Particle-size Analysis. In: Klute, A. (Ed.), METHODS OF SOIIL ANALYSIS: Part 1-Physical and Mineralogical Methods 2nd ed. SSSA, Madison, WI, pp.383-411.

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. GIS 9:421–432.

Gessler, P.E., 1996. Statistical soil-landscape modeling for environmental management. PhD thesis, Australian National University.

Gessler, P.E., O.A. Chadwick, F. Chamran, L. Althouse, and K. Holmes, 2000. Modeling Soil-Landscape and Ecosystem Properties Using Terrain Attributes. Soil Sci. Soc. Am. J. 64:2046-2056.

GRASS Development Team, 2012. Geographic Resource Analysis Support System (GRASS) software. Open source geospatial foundation project. http://grass.osgeo.org

Grunwald, S., 2006. What Do We Really Know about the Space-Time Continuum of Soil-Landscapes? In: Grunwald, S. (Ed.), Environmental Soil-Landscape Modeling – Geographic Information Technologies and Pedometrics. CRC, Boca Raton, FL, pp.3-36.

Guisan, A., T. C. Edwards, Jr., and T. Hastie, 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157:89-100.

Hastie, T., R. Tibshirani, and J. Friedman, 2009. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, NY.

Hengl, T., D.G. Rossiter, and A. Stein, 2004a. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Aust. J. Soil Res. 41(8):1403-1422.

Hengl, T., S. Gruber, and D.P. Shrestha, 2004b. Reduction of errors in digital terrain parameters used in soil-landscape modelling. International Journal of Applied Earth Observation and Geoinformation (JAG), 5(2):97-112.

Hengl, T., 2006. Finding the right pixel size. Computers & Geosciences. 32(9): 1283-1298.

Hengl, T., 2009. A practical guide to geostatistical mapping. University of Amsterdam, Amsterdam, NL, pp.291.

Hengl, T., and R.A. MacMillan, 2009. Geomorphometry – a key to landscape mapping and modeling. In: T. Hengl and H.I. Reuter (Eds.), Geomorphometry: concepts, software and applications. Elsevier, Amsterdam, NL, pp.433-460.

Heuvelink, G.B.M., and R. Webster, 2001. Modeling soil variation: past, present, and future. Geoderma 100:269-301.

Hewitt, A.E., 1993. Predictive modeling in soil survey. Soils and Fertilizers 56:305-314.

Howell, D., Y., Kim, C. Haydu-Houdeshell, P. Clemmer, and R. Almaraz, 2004. Soil Property Distribution Models from Point Data for Soil Survey. 24th Annual ESRI International User Conference, August 9–13, 2004.

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

Jenkins, A.B., 2002. Organic Carbon and Fertility of on the Allegheny Plateau of West Virginia. Masters thesis, West Virginia University.

Jenny, H., 1941. Factors of soil formation: a system of quantitative pedology. McGraw-Hill, New York.

Jenny, H., 1980. The soil resource: origin and behavior. Springer-Verlag. New York.

Lane, P.W., 2002. Generalized linear models in soil science. European Journal of Soil Science 53:241-251.

Liaw, A., and M. Wiener, 2002. Classification and regression by randomForest. R News. 2(3)18-22.

Lierop, W. Van, 1990. Soil pH and Lime Requirement Determination. In: Westerman, R.L. (Ed.), SSSA Book Series 3: Soil Testing and Plant Analysis 3rd ed. SSSA, Madison, WI, pp.73-126.

Lin, H., D. Wheeler, J. Bell, and L. Wilding, 2005. Assessment of soil spatial variability at multiple scales. Ecological Modelling 182:271-290.

MacMillan, R.A., T.C. Martin, T.J. Earle, and D.H. McNabb, 2003. Automated analysis and classification of land-forms using high-resolution digital elevation data: applications and issues. Can. J. Remote Sensing 29(5):592-606.

MacMillan, R.A., W.W. Pettapiece, and J.A. Brierley, 2005. An expert system for allocating soils to landforms through the application of soil survey tacit knowledge. Can. J. Soil Sci. 85:103-112.

MacMillan, R.A., D.E. Moon, R.A. Coupe, and N. Phillips, 2010. Predictive ecosystem mapping (PEM) for 8.2 million ha of Forestland, British Columbia, Canada. J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink, and S. Kienast-Brown (Eds.), Digital soil mapping: bridging research, environmental application, and operation. Springer, New York, NY, pp.337-356.

Grunwald, S., 2006. What Do We Really Know about the Space-Time Continuum of Soil-Landscapes? In: Grunwald, S. (Ed.), Environmental Soil-Landscape Modeling – Geographic Information Technologies and Pedometrics. CRC, Boca Raton, FL, pp.3-36.

Malone, B.P., A.B. McBratney, B. Minasny, and G.M. Laslett, 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154:138-152.

Mechlich, A., 1953. Determination of P, Ca, Mg, K, Na, and NH4. North Carolina Soil Test Division (Mimeo 1953). North Carolina Dep. of Aric., Raleigh, NC.

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

McBratney, A.B., M.L. Mendonca Santos, and B. Minasny, 2003. On digital soil mapping. Geoderma 117: 3-52.

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

McKenzie, N.J., and D.W. Jacquier, 1997. Improving the field estimation of saturated hydraulic conductivity in soil survey. Aust. J. Soil Res. 35:803-826.

McKenzie, N.J., and P.J. Ryan, 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89:67-94.

McKenzie, N.J., Gessler, P.E., Ryan, P.J. and O'Connell, D.A., 2000. The role of terrain analysis in soil mapping. In: J.P. Wilson and J.C. Gallant (Eds.), Terrain Analysis: Principles and Applications. Jon Wiley & Son, New York. pp.245-266.

McKenzie, N.J., A.J. Ringrose-Voase, and M.J. Grundy, 2008. Rationale. In: N.J. McKenzie, M.J. Grundy, R. Webster, and A.J Ringrose-Voase (Eds.), Guidelines for Surveying Soil and Land Resources: Volumen 2 Australian Soil and Land Handbook Series. CSIRO Publishing, Melbourne, pp.1-12.

Minasny, B., and A.B. McBratney, 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32:1378-1388.

Minasny, B., and A.B. McBratney, 2007. Spatial prediction of soil properties using EBLUP with the Matern covariance fuction. Geoderma 140:324-336.

O'Connell, D.A., P.J. Ryan, N.J. McKenzie, and A.J. Ringrose-Voase, 2000. Quantitative site and soil descriptors to improve the utility of forest soil surveys. Forest Ecology and Management 138:107-122.

Odgers, N.P., Z. Libohova, and J.A. Thompson, 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at the continental scale. Geoderma 190:153-163.

Park, S.J., K. McSweeney, and B. Lowery. 2001. Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma 103:249-272.

Park, S.J., and P.L.G. Vlek, 2002. Environmental correlation of three-dimensional soil spatial variability: a comparison of three adaptive techniques. Geoderma 109:117-140.

Park, S.J., and T.P. Burt, 2002. Identification and characterization of pedogeomorphological processes on a hillslope. Soil Sci. Soc. Am. J. 66:1897-1910.

Pennock, D.J., B.J. Zebarth, and E. DE. Jong, 1987. Landform Classification and Soil Distribution in Hummocky Terrain, Saskatchewan, Canada. Geoderma 40:297-315.

PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, created 5 Nov 2011.

R Core Team, 2013. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. URL http://www.R-project.org

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

Ryan, P.J., N.J. McKenzie, D. O'Connell, A.N. Loughhead, P.M. Leppert, D. Jacquier, L. Ashton, 2000. Integrating forest soils information across scales: spatial prediction of soil properties under Australian forests. Forest Ecology and Management 138:139-157.

SAGA GIS Development Team, 2012. System for Automated Geoscientific Analyses (SAGA) software. SAGA user group association. http://www.saga-gis.org

Schmidt, J., I.S. Evans, and J. Brinkmann, 2003. Comparison of polynomial models for land surface curvature. Int. J. GIS 17(8):797-814.

Schmidt J. and A. Hewitt, 2004. Fuzzy land element classification form DTMs based on geometry and terrain position. Geoderma 121:243-256.

Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., and Broderson, W.D. (Eds.), 2002. Field book for describing and sampling soils, Version 2.0. Natural Resource Conservation Service, National Soil Survey Center, Lincoln, NE.

Scull, P., G. Okin, O.A. Chadwick, and J. Franklin, 2005b. A Comparison of Methods to Predict Soil Surface Texture in an Alluvial Basin. The Professional Geographer 57(3)423-437.

Shi, X., A.X. Zhu, J. Burt, W. Chol, R. Wang, T. Pel, B. Li, and C. Qin, 2007. An Experiment Using a Circular Neighborhood to Calculate Slope Gradient from a DEM. Photogrammetric Engineering & Remote Sensing 73(2):143-154.

Smith, M.P., A.X. Zhu, J.E. Burt, and C. Stiles, 2006. The Effects of DEM Resolution and Neighborhood Size on Digital Soil Survey. Geoderma 137:58-69.

Sponaugle, C.L., 2005. Properties and Acid Risk Assessment of Soils in Two Parts of the Cherry River Watershed, West Virginia. Masters thesis, West Virginia University.

Sylvian, J-D., A.R. Michaud, M.C. Nolin, and G.B. Benie, 2012. A novel spectro-temporal approach for predicting soil physical properties. In: B. Minasny, B.P. Malone, and A.B. McBratney (Eds.), Digital soil assessment and beyond: proceedings of the fifth global workshop on digital soil mapping. CRC Press, AK Leiden, pp.380-386.

Tarboton, D. 2004. Terrain using digital elevation models (TauDEM). Available at hydrology.neng.usu.edu/taudem/. Utah State University.

Therneau, T., B. Atkinson, and B. Ripley, 2013. rpart: recursive partitioning. R package version 4.1-1.

Thompson, J.A., J.C. Bell, and C.A. Butler, 1997. Quantitative soil-landscape modeling for estimating the areal extent of hydromorphic soils. Soil Sci. Soc. Am. J. 61:971-980.

Thompson, J.A., J.C. Bell, and C.A. Bulter, 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma 100:67-89.

Thompson, J.A., E.M. Pena-Yewtukhiw, and J.H. Grove, 2006. Soil-landscape modeling across a physiographic region: Topographic patterns and model transportability. Geoderma 133:57-70.

United States Department of Agricultural, Natural Resources Conservation Service. 2006. Land Resource Regions and Major Land Resource Areas of the United States, the Caribbean, and the Pacific Basin. U.S. Department of Agricultural Handbook 296.

United States Department of Interior, U.S. Geological Survey, 2006. Multi-resolution Land Characteristics 2001 Image Processing Procedure. White Paper pp.1-9.

Webster, R., and M.A. Oliver 1990. Statistical Methods in Soil and Land Resource Survey. Oxford University Press, Oxford.

Webster, R., 2000. Is soil variation random? Geoderma 97:149-163.

Webster, R., 2001. Statistics to support soil research and their presentation. European Journal of Soil Science 52:331-340.

Webster, R., and M.A. Oliver 2001. Geostatistics for Environmental Scientists. Wiley & Sons, Chichester.

West Virginia Geological and Economic Survey, 1968. 1968 Geological Map of West Virginia.William & Heintz Map Corporation, Washington D.C., VA.

Wilson, J.P., P.L. Repetto, and R.D. Snyder, 2000. Effect of Data Source, Grid Resolution, and Flow-Routing Method on Computed Topographic Attributes. In: J.P. Wilson and J.C. Gallant (Eds.) Terrain analysis: Principles and Applications. John Wiley&Sons, New York, pp.133–161.

Wood, J., 1996. The geomorphological characterization of Digital Elevation Models. PhD Thesis Department of Geography, University of Lancaster, UK.

Yemefack, M., D.G. Rossiter, and R. Njomgang, 2005. Multi-scale characterization of soil variability within an agricultural landscape mosaic system in southern Cameroon. Geoderma 125:117-143.

Young, M., 1978. Terrain analysis: program documentation. Report 6 on Grant DA-ERO-591-73-G0040. Statistical characterization of altitude matrices by computer. Department of Geography, University of Durham, Durham, UK, pp.1-27.

Young, F. J., and R. D. Hammer, 2000. Defining Geographic Soil Bodies by Landscape Position, Soil Taxonomy, and Cluster Analysis. Soil Sci. Soc. Am. J. 64:989-998.

Zhu, A.X., and L.E. Band, 1994. A Knowledge-Based Approach to Data Integration for Soil Mapping. Canadian Journal of Remote Sensing 20(4):408-418.

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

Zhu, A.X., L. Band, R. Vertessy, and B. Dutton, 1997. Derivation of soil properties using a soil land inference model (SoLIM). Soil Sci. Soc. Am. J. 61:523-533.

Zhu, A.X., 1997b. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. Am. Soc. Photo and Rem. Sensing 63: 1195-1202.

Zhu, A.X., 2006. Fuzzy Logic Models. In: Grunwald, S. (Ed.), Environmental Soil-Landscape Modeling – Geographic Information Technologies and Pedometrics. CRC, Boca Raton, FL, pp. 215–239.

Appendix

Appendix 1: Soil dataset

pedon id	horizon	suffix	top (cm)	bottom (cm)	fragvol (% volume)	clay (% weight)	sand (% weight)	pH	bulk density (g/cm^3)	C (kg/ha)	P (kg/ha)	Al (kg/ha)	Ca (kg/ha)	Mg (kg/ha)
3	A	NA	0	15	45	7	72	3.19	1.121	3.68	5.55	548.84	35.64	9.86
3	Bw	NA	15	31	40	13	62	4.03	1.4	2.39	105.37	2040.88	42.82	6.94
3	C	NA	31	48	45	17	64	4.04	1.589	1.25	6.45	1388.42	46.71	6.48
8	A	NA	0	20	25	22	46	3.86	1.056	7.28	8.79	181.41	98.31	20.03
8	Bw	1	20	44	10	4	60	4.19	1.488	3.22	27.07	4442.37	98.73	13.07
8	Bw	2	44	65	8	5	69	4.17	1.682	1.82	41.31	3949.01	115.34	14.15
8	C	NA	65	92	45	1	66	4.26	1.868	0.99	19.57	2211.84	87.79	9.9
10	A	NA	0	11	16	28	45	4.03	1.093	3.92	1.46	1259.15	26.28	7.48
10	AB	NA	11	29	4	26	42	4.18	1.36	4.11	0.12	3310.19	83.28	12.49
10	Bt	1	29	44	7	23	50	3.98	1.554	1.63	-0.52	1558.32	22.89	4.88
10	Bt	2	44	71	8	8	67	3.99	1.719	2.29	2.8	2295.05	40.35	7.05
10	Bt	3	71	99	35	16	53	4.14	1.914	1.86	1.29	1809.99	38.41	6.62
10	Bt	4	99	119	2	25	29	3.81	2.084	2.23	-2.43	2435.93	139.44	20.72
10	Bt	5	119	150	7	22	52	4.04	2.265	3.5	3.56	4483.31	106.91	15.38
11	A	NA	0	14	25	24	91	3.02	1.255	2.28	4.85	31.53	105.15	14.53
11	C	NA	14	30	20	4	92	3.12	1.455	0.95	1.61	31.45	70.14	8.19
12	A	NA	0	17	5	16	48	2.84	0.798	9.11	1.84	939.74	73.23	19.68
12	BE	NA	17	48	4	15	49	3.73	1.234	8.3	0.88	4105.17	125.39	25.74
12	Bt	1	48	74	30	16	52	4.01	1.401	2.84	4.53	4046.39	82.37	10.78
12	Bt	2	74	95	22	16	49	3.96	1.511	1.33	1.23	2699.46	85.02	11.91
12	Bt	3	95	109	5	37	33	3.92	1.549	1.4	0.63	1528.01	88.94	11.24
12	Btg	NA	109	150	2	40	17	3.76	1.638	3.45	2.04	4867.29	223.82	32.05
13	A	NA	0	7	5	28	35	3.09	1.072	2.34	1.14	485.25	40.22	9.21
13	Bt	1	7	39	5	29	33	3.9	1.225	7.44	3.61	2094.26	136.75	21.85
13	Bt	2	39	57	14	24	37	3.88	1.386	1.74	1.51	1191.48	77.55	12.09
13	Bt	3	57	150	52	15	55	3.91	1.54	5.67	3.16	6741.76	71.58	30.93
18	A	NA	0	7	65	15	75	3.03	1.081	0.88	1.01	67.31	24.09	5.52
18	E	NA	7	28	45	6	74	4.06	1.332	0.9	1	230.77	66.96	13.19
18	Bt	1	28	77	20	20	42	3.54	1.404	5.07	7.07	2503.73	297.7	115.86
18	Bt	2	77	101	40	9	56	3.7	1.529	1.35	0.67	1427.05	136.21	44.62
18	Bt	3	101	150	55	16	54	3.99	1.62	2.7	3.36	2352.47	172.3	54.61
21	A	NA	0	12	11	34	24	3.95	1.075	4.2	0.76	1027.11	26.91	15.87
21	Bw	1	12	35	30	28	30	3.96	1.162	5.92	1.93	3195.63	24.32	9.58
21	Bw	2	35	44	7	22	23	4.45	1.364	1.27	0.33	1176.83	14.87	3.6
21	Bw	3	44	59	10	21	30	4.2	1.313	3.91	1.99	3467.5	29.43	4.7
21	BC	NA	59	150	45	5	67	3.99	1.55	8.12	12.69	6937.47	96.57	17.65
24	Bt	1	0	30	44	16	35	3.72	1.314	1.59	0.85	1935.35	35.53	7.85
24	Bt	2	30	53	30	24	26	3.87	1.388	1.61	0.11	2188.44	45.04	7.98
24	Bt	3	53	67	5	17	47	3.91	1.383	2.7	1.72	2017.34	37.19	8.8
25	A	NA	0	10	18	35	8	3.68	0.885	4.21	0.44	982.17	124.99	22.44
25	Bt	NA	10	28	25	26	24	3.97	1.208	3.43	1.59	1923.98	146.72	25.56
25	BC	NA	28	40	45	19	29	3.94	1.301	1.24	0.46	1063.45	70.41	10.14
25	C	NA	40	58	10	26	32	3.87	1.295	4.33	0.05	2403.4	478.48	65.74
26	A	NA	0	10	0	7	7	3.04	0.802	5.61	2.39	465.33	151.51	26.97
26	Bt	1	10	20	3	34	8	3.51	1.242	1.95	0.44	1273.89	60.8	11.99
26	Bt	2	20	33	5	36	9	3.74	1.288	2.3	0.95	1976.77	65.6	13.36
26	Bt	3	33	47	0	30	9	3.79	1.375	1.54	0.66	2262.43	77.85	15.11
26	Bt	4	47	64	8	28	8	3.78	1.437	1.26	0.34	2074.59	113.84	23.25
28	A	NA	0	17	0	30	14	3.44	1.232	2.8	2.68	3454.28	80.97	21.12
28	Bt	1	17	31	10	28	18	3.82	1.308	1.5	1.36	1816.52	99.73	24.03
28	Bt	2	31	47	10	19	30	3.82	1.373	1.22	-1.04	2146.78	98.62	23.37
28	Bt	3	47	76	5	21	22	3.84	1.444	2.03	1.75	2523.69	503.35	78.67
28	Bt	4	76	102	0	24	11	3.97	1.522	1.96	2.27	2858.14	1612.63	572.02
28	Bt	5	102	150	4	29	3	3.92	1.63	3.26	8.41	4625.58	4929.69	2227.94
30	A	NA	0	5	20	4	17	3.14	0.653	2.25	0.47	202.55	54.45	8.84
30	BA	NA	5	21	0	44	11	3.72	0.944	7.88	2.13	2456.22	50.94	17.93
30	Bt	1	21	42	10	39	39	4.12	1.269	3.98	-0.7	5083.38	30.72	7.27
30	Bt	2	42	69	8	25	26	3.93	1.418	2.07	-1.13	3651.24	44.09	11.68
30	Bt	3	69	86	8	15	44	3.91	1.492	0.98	3.55	1624.52	36.7	9.23
31	A	NA	0	16	8	42	27	3.56	0.895	7.66	0.85	1507.81	49.88	10.96
31	EB	NA	16	38	10	24	33	4.17	1.251	4.64	1.59	4511.51	80.98	10.98
31	BE	NA	38	53	7	24	28	4.18	1.286	3.75	1.45	3249.93	27.77	4.31
31	Bt	NA	53	90	10	16	71	4.24	1.44	4.96	4.31	7639.19	153.81	20.09
31	BC	NA	90	102	2	9	47	3.88	1.547	0.96	0.97	1182.68	30.13	6.42
32	A	NA	0	12	20	12	53	3.16	0.969	4.52	0.87	990.52	20.39	7.21
32	Bw	NA	12	39	40	11	54	3.94	1.3	2.64	0.71	2055.22	64.31	11.43
33	A	NA	0	32	17	17	63	4.07	1.278	3.7	6.24	4789.95	70.48	11.91
33	Bw	1	32	63	10	15	62	4.16	1.351	4.83	6.77	4613.06	59.6	10.23

pedon id	horizon	suffix	top (cm)	bottom (cm)	fragvol (% volume)	clay (% weight)	sand (% weight)	pH	bulk density (g/cm^3)	C (kg/ha)	P (kg/ha)	Al (kg/ha)	Ca (kg/ha)	Mg (kg/ha)
33	Bw	2	63	107	40	16	68	4.28	1.511	2.23	6.72	4045.14	53.9	9.49
33	Bw	3	107	150	30	8	76	3.96	1.637	2.46	7.24	2653.55	175.82	22.91
34	A	1	0	17	30	21	56	3.73	1.087	4.22	3.16	1692.69	125.53	17.52
34	A	2	17	35	17	15	55	4.08	1.231	3.72	4.41	2643.01	77.6	10.08
34	BA	NA	35	69	8	15	52	4.3	1.37	4.98	5.83	7091.34	73.91	7.48
34	Bt	1	69	100	8	16	46	3.99	1.513	2.08	3.9	2213.39	420.03	55.72
34	Bt	2	100	150	0	20	20	3.69	1.633	3.15	4.16	5381.69	566.34	146.34
40	A	NA	0	9	2	4	20	3.08	0.819	4.75	0.78	409.35	41.76	10.33
40	BA	NA	9	28	2	24	26	3.68	1.256	3.26	-0.57	1777.15	30.2	9.48
40	Bt	NA	28	46	5	16	53	3.96	1.33	2.49	1.62	1709.49	31.38	6.8
43	A	NA	0	8	0	39	14	3.48	0.921	3.85	0.9	619.79	36.9	13.67
43	BA	NA	8	24	2	39	13	3.78	1.249	2.72	0.02	1559.35	34.22	9.44
43	Bt	1	24	55	6	32	18	4	1.324	4.89	-0.85	4381.4	126.16	21.01
43	Bt	2	55	91	8	24	30	4.07	1.451	4	-1.2	4801.28	193.2	26.91
43	Bt	3	91	150	5	32	22	4.02	1.586	7.17	0.18	10935.9	272.93	45.1
44	A	NA	0	10	0	19	44	3.62	1.152	2.72	0.62	1060.75	44.92	11.37
44	Bt	NA	10	25	20	20	50	3.69	1.25	2.3	0.29	1182.38	17.02	5.95
44	BC	NA	25	50	12	15	50	3.71	1.285	5.11	0.99	2485.64	34.53	9.74
46	A	NA	0	8	2	8	17	3.22	1.017	3.23	0.8	436.88	64.77	11.21
46	Bt	1	8	27	2	29	16	3.71	1.211	4.55	-0.99	2102.03	97.99	22.08
46	Bt	2	27	49	5	31	17	3.92	1.328	3.44	0.61	2377.74	56.32	10.74
46	Bt	3	49	89	20	21	40	4.12	1.419	5.36	2.61	5570.51	102.08	17
46	Bt	4	89	135	75	28	37	3.98	1.533	2.21	-0.75	1547.3	78.05	13.43
48	A	NA	0	10	0	31	6	3.43	1.067	3.7	1.18	985.48	29.13	11.99
48	BA	NA	10	38	0	38	14	3.44	1.26	5.82	-1.94	3648.96	138.5	35.4
48	Btg	1	38	58	3	25	22	3.7	1.372	2.83	-1.24	2262.36	92.73	19.9
48	Btg	2	58	74	10	16	36	3.87	1.437	1.8	0.65	1618.35	28.03	7.96
48	Btg	3	74	95	10	21	31	3.76	1.494	2.25	0.4	2013.5	100.27	18.19
50	C	NA	0	8	85	4	77	3.12	1.297	0.14	1.53	31.04	2.58	0.49
52	A	NA	0	7	0	17	17	2.9	0.918	3.4	0.74	242.15	30.21	8.71
52	BA	NA	7	19	0	40	13	3.32	1.213	2.6	-0.25	1169.06	60.13	13.77
52	Bt	NA	19	34	30	33	45	3.57	1.286	1.8	-0.64	904.5	47.97	10.06
52	BC	NA	34	56	17	21	37	3.83	1.353	2.79	2.11	1679.76	62.72	9.09
54	Oa	1	0	61	90	NA	NA	NA	NA	NA	NA	NA	NA	NA
54	Oa	2	61	150	90	NA	NA	NA	NA	NA	NA	NA	NA	NA
59	AE	NA	0	8	3	9	69	3.11	1.141	2.31	1.67	435.15	18.81	6.62
59	BE	NA	8	18	5	11	60	3.07	1.273	1.48	-0.07	554.75	41.47	6.39
59	Bss	NA	18	40	7	14	68	4.03	1.361	1.83	21.21	2931.03	96.13	13.31
60	A	1	0	14	10	35	34	3.81	1.143	3.78	2.16	939.9	441.72	55.04
60	A	2	14	35	10	33	33	3.86	1.238	4.74	4.69	2075.53	394.28	52.88
60	BA	NA	35	46	4	23	33	3.84	1.324	2.08	-0.81	1164.35	207.22	44.94
60	Bt	1	46	59	8	26	35	4.01	1.404	1.48	2.38	1089.89	278.02	41.87
60	Bt	2	59	81	8	22	39	3.82	1.464	2.17	1.35	1634.59	437.92	69.67
60	Bt	3	81	150	30	23	42	3.82	1.607	4.11	1.71	4191.09	1351.24	268.81
61	A	NA	0	25	50	28	27	3.73	0.991	5.82	4.11	1329.46	209.21	55.85
63	C	2	0	55	90	7	75	2.86	1.332	1.01	3.96	151.93	25.54	3.67
63	C	3	55	99	85	12	67	2.91	1.492	1.08	8.87	359.36	18.65	2.83
68	A	NA	0	10	3	32	15	3.1	0.78	5.37	1.12	795.79	39.22	17.06
68	Bt	1	10	32	3	28	20	3.71	1.179	6.26	0.03	3385.83	34.2	16.02
68	Bt	2	32	54	3	22	39	4.03	1.375	2.16	1.22	3707.83	92.93	20.39
68	Bt	3	54	108	50	14	39	3.85	1.502	1.89	44.8	3135.83	129.95	29.68
68	Bt	4	108	150	65	22	39	3.74	1.642	0.91	21.97	1962.66	111.7	25.54
69	A	NA	0	12	5	13	38	3.22	0.859	5.98	0.86	853.83	108.31	18.95
69	BA	NA	12	37	8	26	37	4.08	1.18	7.04	3.14	4831.95	55.87	9.58
69	Bw	NA	37	57	17	21	43	4.3	1.345	2.84	2.78	2948.77	28.32	6.29
69	Bt	1	57	87	25	24	40	3.91	1.476	1.55	1.28	2007.89	40.17	10.53
69	Bt	2	87	150	20	30	34	3.89	1.612	3.15	3.25	6614.2	197.33	45.91
72	A	NA	0	21	5	27	31	4.14	1.17	5.1	3.55	3085.26	40.93	11.18
72	BA	NA	21	42	5	27	33	4.04	1.343	1.84	0.8	2587.75	34.48	7.42
72	Bt	1	42	61	5	22	71	3.88	1.416	1.2	-1.05	1584.65	28.14	7.23
72	Bt	2	61	95	13	24	35	3.83	1.491	1.99	2.29	3121.19	60.33	22.79
72	Bt	3	95	150	30	30	31	3.85	1.618	2.63	0.5	3927.49	249.42	119.1
74	A	NA	0	21	25	23	31	3.75	0.925	7.94	-0.04	1599.77	102.23	26.32
74	Bt	1	21	57	12	30	37	4.04	1.352	3.93	1.8	5970.18	182.94	30.14
74	Bt	2	57	74	5	22	40	3.96	1.443	1.63	0.85	2050.82	95.74	19.3
74	Bt	3	74	92	9	13	56	4.04	1.507	1.24	3.07	1635.52	88.48	12.27
74	Btx	NA	92	150	12	15	53	4.01	1.61	4.49	9.82	6618.43	335.59	50.61
79	A	NA	0	23	5	28	25	3.57	1.113	7.02	1.52	2155.25	118.36	29.14

pedon id	horizon	suffix	top (cm)	bottom (cm)	fragvol (% volume)	clay (% weight)	sand (% weight)	pH	bulk density (g/cm^3)	C (kg/ha)	P (kg/ha)	Al (kg/ha)	Ca (kg/ha)	Mg (kg/ha)
79	BA	NA	23	52	7	18	34	3.96	1.336	3.33	-0.99	2571.73	114.82	19.28
79	Bt	1	52	84	18	13	48	3.9	1.444	2.4	-1.04	2703.54	120.7	20.27
79	Bt	2	84	125	40	22	47	3.83	1.551	2.18	-0.8	2231.66	112.19	19.26
79	Bt	3	125	150	12	16	60	3.93	1.659	1.29	-0.55	1399.37	114.34	17.16
80	A	NA	0	16	45	24	40	3.34	1.19	2.03	0.82	583.19	55.49	9.59
80	BA	NA	16	28	23	31	32	3.57	1.267	1.69	0.46	977.18	47.38	8.94
80	Bt	NA	28	56	40	29	32	3.91	1.319	3.33	-0.14	2310.51	87.36	13.66
80	Btx	NA	56	101	55	14	48	3.91	1.488	1.86	0.92	1882.3	112.65	16.05
81	A	NA	0	31	17	30	26	4.13	1.249	4.42	7.39	3559.8	124.53	16.11
81	Bw	NA	31	85	17	22	23	3.94	1.415	4.77	4.88	6708.38	496.71	67.34
81	Bt	NA	85	150	75	19	46	4	1.581	1.94	1.19	2352.55	211.61	30.82
84	A	NA	0	13	6	42	13	3.69	1.194	2.83	2.03	1409.81	89.03	20.21
84	Bt	1	13	34	10	28	29	3.69	1.329	2	0.95	1915.95	239.09	34.03
84	Bt	2	34	49	30	18	42	3.89	1.39	0.96	0.64	1010.14	190.06	21.69
85	A	NA	0	18	20	23	35	4.16	1.172	4.17	5.09	1409.49	826.8	61.22
85	Bt	1	18	41	30	20	35	3.94	1.361	1.63	1.45	1169.36	292.72	34.54
85	Bt	2	41	54	12	26	32	3.84	1.423	0.97	0.64	788.39	227.58	26.92
85	Bt	3	54	76	27	19	41	3.98	1.477	1.26	0.88	1168.76	455.53	82.58
85	Bt	4	76	107	40	15	45	3.96	1.549	1.61	1.48	1213.02	456.63	103.7
86	A	NA	0	15	25	40	37	3.57	1.078	4.01	3.48	798.55	202.4	23.08
86	Bt	1	15	40	25	19	37	3.94	1.293	2.98	3.65	1855.34	146.97	16.48
86	Bt	2	40	71	40	23	47	3.94	1.392	2.48	8	1579.86	203.7	22.09
86	C	1	71	114	55	19	52	3.99	1.501	2.57	2.51	2096.18	296.34	38.15
86	C	2	114	150	20	24	32	3.79	1.608	4.33	10.03	2542.39	326.25	39.46
87	A	NA	0	10	60	15	50	3.22	1.087	1.38	0.78	242.35	26.37	9.07
87	Bt	1	10	28	35	17	32	4.2	1.25	2.28	3.96	1629.2	21.37	7.17
87	Bt	2	28	46	25	17	37	4.08	1.364	1.33	1.67	1223.1	74.05	17.51
87	Bt	3	46	66	20	21	13	3.91	1.431	1.24	-1	1096.12	37.4	47.5
89	A	NA	0	15	45	6	66	3.33	0.803	4.61	1.67	430.47	61.33	10.94
89	BA	NA	15	28	20	9	62	4.21	1.215	2.66	2.24	1842.3	46.35	6.9
89	Bt	1	28	64	30	15	57	4.14	1.37	3.36	2.37	2125.43	125.63	17.9
89	Bt	2	64	94	40	16	50	3.84	1.503	1.2	-1.52	1647.87	96.07	16.03
89	Bt	3	94	150	80	18	54	4.02	1.615	1.01	0.57	1237.8	70.14	11.3
90	C	NA	0	8	22	19	40	3.68	1.267	0.77	-0.01	416.17	9.66	3.19
90	Bt	1	8	38	30	18	37	3.97	1.349	1.68	-1.22	1352.8	94.56	22.06
90	Bt	2	38	71	25	25	33	3.91	1.429	2.49	1.7	2335.61	51.87	13.45
90	Bt	3	71	150	40	15	37	3.68	1.564	7.26	6.45	5660.19	116.98	28.43
92	A	NA	0	14	30	31	29	3.35	1.01	4.21	0.61	830.1	54.07	16.98
92	BA	NA	14	30	17	37	19	3.86	1.222	3.31	1.36	1771.41	59.27	14.73
92	Bw	1	30	49	30	16	39	3.83	1.355	1.69	0.69	1264.36	23.43	10.61
92	Bw	2	49	62	8	24	24	3.84	1.403	1.49	0.86	1010.36	110.44	59.71
92	Bw	3	62	83	5	19	27	3.79	1.467	1.98	1.1	1205.09	224.22	194.65
92	Bg	NA	83	110	5	23	15	3.82	1.539	2.44	2.03	1337.02	344.96	321.67
92	BC	NA	110	150	25	20	23	3.87	1.627	3.34	2.37	1545.01	308.03	416.4
93	A	NA	0	20	30	20	39	3.24	1.13	4.6	3.29	714.67	98.75	23.72
93	Bt	1	20	46	60	19	36	3.94	1.236	3.03	3.62	1381.39	90.89	13.05
93	Bt	2	46	62	62	23	38	4.14	1.334	1.5	1.74	1065.02	34.53	5.72
93	Bt	3	62	137	40	15	44	3.7	1.547	5.1	-3.1	3514.65	366.43	149.6
94	A	NA	0	16	42	25	37	3.51	0.812	5.23	1.06	965.1	25.66	8.68
94	BA	NA	16	29	11	17	37	4.14	1.059	5.21	3.29	3058.59	21.28	5.15
94	Bw	1	29	50	16	15	51	4.22	1.232	5.67	7.71	5381.25	50.44	7.27
94	Bw	2	50	73	18	14	58	4.18	1.399	3.25	8.31	4421.79	40.6	6.12
94	Bw	3	73	93	25	16	63	4.17	1.497	1.67	4.5	2594.53	24.54	3.91
94	BC	NA	93	150	70	10	58	4.11	1.578	2.93	3.32	1968.65	59.99	8.37
96	A	NA	0	12	5	42	12	3.51	0.899	5.68	0.94	1075.01	56.9	20
96	E	NA	12	34	20	23	18	3.81	1.169	5.46	1.66	1969.53	104.8	21.98
96	Bt	1	34	77	40	24	27	3.94	1.265	8.44	-0.96	3290.63	45.16	16.12
96	Bt	2	77	103	40	20	37	3.93	1.436	3.62	1.21	1908.15	91.84	15.51
96	Bt	3	103	150	40	14	47	3.84	1.555	6.11	3.44	2849.6	185.49	32.88
97	A	NA	0	12	40	26	27	4.27	1.088	2.45	2.26	535.42	621.98	45.47
97	Bt	NA	12	40	60	22	29	4.16	1.316	1.3	1.44	1087.89	472.31	52.27
101	A	NA	0	15	17	32	15	3.73	1.157	3.28	1.55	1391.35	379.36	58.68
101	Bt	1	15	38	33	22	31	3.82	1.335	1.35	1.76	1179.61	913.45	422.27
101	Bt	2	38	59	13	34	13	3.97	1.398	1.63	1.21	1966.17	681.39	116.53
101	Bt	3	59	108	23	20	44	3.94	1.499	3.33	3.82	4165.53	1008.72	170.58
101	Btg	NA	108	150	37	20	30	3.94	1.626	2.55	3.94	1726.84	3820.22	1031.2
103	A	NA	0	11	55	9	76	4.07	1.3	0.92	2.45	247.11	160.69	27.76
103	Bx	1	11	70	50	7	65	4.51	1.465	2.44	10.05	440.47	1750.44	483.46

pedon id	horizon	suffix	top (cm)	bottom (cm)	fragvol (% volume)	clay (% weight)	sand (% weight)	pH	bulk density (g/cm^3)	C (kg/ha)	P (kg/ha)	Al (kg/ha)	Ca (kg/ha)	Mg (kg/ha)
103	Bx	2	70	150	50	7	66	4.61	1.668	2.99	8.71	375.22	2958.08	840.5
104	A	NA	0	11	0	15	48	3.31	1.059	4.31	2.23	825.68	48.8	14.38
104	E	NA	11	29	7	21	46	3.52	1.27	3.14	-0.62	2080.57	25.61	15.19
104	Bt	1	29	57	5	26	43	3.86	1.351	4.55	0.57	2653.36	125.37	19.69
104	Bt	2	57	93	50	20	47	3.88	1.476	2.14	0.78	1747.61	108.42	18.49
105	A	NA	0	23	10	29	26	3.71	1.036	8.58	4.54	1433.67	774.31	100.1
105	Bt	1	23	60	12	18	34	3.74	1.276	8.17	5.82	3656.18	536.79	71.5
105	Bt	2	60	94	12	16	38	3.89	1.424	5.67	4.13	2786.42	396.66	67.79
107	A	NA	0	24	0	8	69	3.4	1.288	4.53	5.1	1348.4	113.91	20.48
107	Bg	1	24	65	0	28	22	3.74	1.383	7.99	31.98	3966.45	102.61	24.81
107	Bg	2	65	132	0	13	57	3.64	1.571	9.68	42.06	3323.8	436.51	258.21
110	A	NA	0	18	5	33	27	3.87	1.109	5.74	4.77	934.22	416.65	100.06
110	Bt	1	18	48	10	24	19	3.88	1.346	2.96	-0.78	1713.29	511.11	246.65
110	Bt	2	48	76	20	22	31	3.96	1.429	2.53	2.08	1146.13	545.53	288.77
110	C	NA	76	97	40	17	31	3.92	1.5	1.44	0.96	555.13	377.9	222.39
111	A	NA	0	10	22	33	24	4.63	1.196	1.74	3.58	795.38	44.07	7.49
111	Bw	NA	10	17	14	10	59	4.21	1.238	1.21	1.67	596.19	30.83	4.65
111	C	NA	17	40	65	10	61	3.96	1.356	0.69	0.22	393.24	37.94	5.67
113	A	NA	0	10	25	10	53	3.07	0.869	4.04	1.19	502.44	31.14	13.63
113	Bw	1	10	23	20	15	47	4.01	1.282	1.65	3.3	1093.08	47.58	10.04
113	Bw	2	23	56	75	17	48	3.97	1.388	0.79	0.61	660.66	36.49	6.83
114	A	NA	0	18	20	7	70	3.28	0.984	6.61	2.87	924.89	121.03	18.81
114	Bw	1	18	33	20	8	66	4.24	1.303	1.73	2.13	1234.16	54.8	7.2
114	Bw	2	33	76	40	11	68	4.03	1.422	2.24	2.27	1924.34	114.71	16.71
121	A	NA	0	19	10	23	29	3.78	1.019	7.41	5.99	848.48	624.14	69.91
121	BA	NA	19	49	30	25	37	3.86	1.283	4.61	3.25	1533.28	428.19	58.68
121	Bt	1	49	84	18	25	32	4.04	1.433	3.89	2.45	1714.07	1274.97	214.82
121	Bt	2	84	107	25	17	40	4.18	1.524	2.13	0.11	1226.93	638.05	151.56
121	Bt	3	107	150	65	11	47	4.23	1.62	1.87	0.49	1078.9	774.71	171.59
123	E	NA	0	10	35	10	61	3.51	1.324	0.64	-0.07	254.96	108.87	9.2
123	Bt	1	10	22	15	18	41	3.96	1.322	1.57	1	1243.88	185.74	16.47
123	Bt	2	22	40	30	13	49	4.04	1.389	1.46	0	1314.75	187.5	17.94
123	Bt	3	40	59	14	15	50	4.08	1.463	1.32	-0.67	1247.64	294.58	31.45
123	Bt	4	59	89	7	21	46	3.93	1.534	2.19	0.6	2455.06	813.14	92.29
123	Bt	5	89	110	5	41	23	3.83	1.606	1.62	0.32	1946.51	1025.8	146.97
123	Bt	6	110	150	7	43	7	3.86	1.704	2.12	-3.42	3448.96	2693.55	516.39
126	A	NA	0	16	35	30	29	3.85	0.972	4.76	2.62	732.14	441.1	45.66
126	BA	NA	16	32	13	26	24	4.36	1.312	1.57	2.04	1106.98	619.98	93.89
126	Bt	1	32	54	35	19	37	4.03	1.376	1.42	0.4	868.67	641.37	100.81
126	Bt	2	54	100	45	17	43	4.09	1.484	2.07	2.61	1676.26	1518.6	263.7
126	C	NA	100	167	0	20	47	4.5	1.635	6.94	48.04	4682.08	9553.71	1525.63
127	A	NA	0	11	0	45	12	3.48	0.932	5.27	1.67	1421.96	45.68	16.04
127	Bt	1	11	25	0	39	13	3.68	1.245	2.65	-0.85	1930.08	97.33	19.75
127	Bt	2	25	36	0	36	26	3.73	1.336	1.16	0.1	1603.14	36	8.35
127	Bt	3	36	89	0	19	43	3.86	1.443	4.27	3.52	7126.78	198.32	39.65
128	A	NA	0	9	3	36	34	3.65	1.129	2.53	1.2	775.04	72.87	12.5
128	BA	NA	9	24	5	33	27	3.74	1.214	3.29	0.5	1574.8	108.27	17.74
128	Bt	1	24	45	3	25	38	4.04	1.317	3.25	2.71	3595.25	68.55	10.56
128	Bt	2	45	74	11	18	35	3.91	1.423	2.73	3.36	3259.73	267.94	36.45
130	A	NA	0	10	0	27	23	3.05	1.115	3.41	-0.1	1032.72	81.19	11.16
130	Bt	1	10	28	0	25	27	3.83	1.258	3.87	0.05	3055.11	42.79	7.39
130	Bt	2	28	38	0	21	33	4.18	1.367	1.09	1.22	2058.53	46.42	5.42
130	BC	NA	38	51	0	17	35	3.96	1.414	1.11	1.06	1863.32	30.25	5.15
130	C	1	51	87	0	19	42	3.68	1.493	2.5	11.48	3467.27	244.13	67.63
130	C	2	87	150	25	24	33	3.81	1.631	3.64	21.5	4373.14	530.56	152.5
136	A	NA	0	18	6	28	34	3.31	0.854	9.28	1.68	2066.72	52.61	18.41
136	C	NA	18	47	77	21	37	3.87	1.03	3.43	1.05	1289.95	35.65	6.81
137	AB	NA	0	29	10	13	58	2.91	1.315	4.39	1.6	1938.68	53.24	8.99
137	Bss	NA	29	52	7	11	41	3.73	1.299	6.75	0.03	3666.79	35.64	9.81
137	Bw	NA	52	74	50	15	49	4.28	1.481	1.44	1.56	3817.7	17.63	1.8
137	BC	NA	74	123	75	9	36	4.05	1.606	1.07	7.56	1851.26	62.56	7.08
138	A	NA	0	16	0	38	25	4.23	0.988	7.17	3.51	659.4	2667.32	251.4
138	Bt	1	16	38	16	24	31	4.06	1.321	2.17	1.72	1228	2065.39	333.49
138	Bt	2	38	57	19	34	33	4.16	1.395	1.43	1.56	1159.62	1346.44	271.11
138	Bt	3	57	96	20	27	33	4.18	1.481	2.82	0.05	2210.12	3365.53	751.87
138	Bt	4	96	150	50	28	42	4.16	1.612	2.57	4.66	1688.06	2535.24	578.21
140	Bt	1	0	23	0	31	24	3.59	1.263	3.56	1.45	3147.54	217	40.5
140	Bt	2	23	47	16	24	26	3.73	1.311	3.84	-0.53	2913.04	71.39	20.33

pedon id	horizon	suffix	top (cm)	bottom (cm)	fragvol (% volume)	clay (% weight)	sand (% weight)	pH	bulk density (g/cm^3)	C (kg/ha)	P (kg/ha)	Al (kg/ha)	Ca (kg/ha)	Mg (kg/ha)
140	Bt	3	47	73	19	16	49	3.83	1.469	0.95	9.85	2080.53	152.71	38.84
140	Bt	4	73	104	20	21	25	3.81	1.532	1.98	4.05	2525.11	213.37	38.08
140	C	NA	104	150	50	17	44	3.78	1.656	1.19	82.2	2498.26	1161.16	356.84
142	A	NA	0	11	8	16	56	3.35	1.03	4.17	1.73	888.3	40.25	15.28
142	AB	NA	11	39	10	22	49	4.01	1.211	7.19	7.17	6128.74	114.85	23.38
142	C	NA	39	60	80	17	44	4.2	1.37	0.68	2.05	1073.36	19.67	3.09
143	A	NA	0	22	4	37	25	4.56	1.028	8.79	8.55	761.33	3788.71	292.43
143	BA	NA	22	40	6	33	29	4.04	1.232	4.42	0.23	1327.8	327.68	45.53
143	Bt	1	40	69	12	27	33	3.75	1.402	2.72	4.56	2225.47	864.97	126.53
143	Bt	2	69	150	20	25	40	3.99	1.558	7.35	13.53	6643.73	6017.73	686.08
144	A	NA	0	15	40	11	62	3.22	1.159	2.69	0.27	403.96	197.99	16.21
144	BA	NA	15	36	10	13	54	4.11	1.227	5.45	3.87	4613.11	96.32	10.75
144	Bt	1	36	63	33	8	61	4.22	1.408	2.19	5.13	3722.32	93.26	11.66
144	Bt	2	63	83	20	12	31	3.99	1.5	1.25	1.43	1736.56	152.65	17.16
146	A	NA	0	21	2	28	34	3.45	0.853	11.21	5.7	2192.89	208.5	34.42
146	BA	NA	21	44	2	26	27	4.06	1.286	4.49	7.24	3753.96	117.88	17.44
146	Bt	1	44	65	2	6	38	4.01	1.416	1.91	4.59	2631.03	115.75	18.22
146	Bt	2	65	80	2	19	34	3.89	1.491	0.71	3.99	1759.44	93.1	15.45
146	Bt	3	80	102	3	15	35	3.82	1.532	1.52	11.35	2310.35	154.4	32.34
146	BC	NA	102	150	2	21	37	3.79	1.642	2.4	32.44	5808.04	511.29	169.84
147	A	NA	0	15	40	26	35	3.87	1.223	2.2	4.23	975.81	165.72	22.64
147	Bt	1	15	35	40	21	34	4.11	1.257	3.31	5.17	2485.81	155.89	18.77
147	Bt	2	35	61	40	28	31	3.98	1.417	2.11	2.67	1703.83	326.18	62.56
147	Bw	1	61	77	50	10	58	3.95	1.522	0.42	6.14	453.02	316.79	78.34
147	Bw	2	77	105	27	9	63	3.91	1.574	1.49	22.78	1232.25	803.64	201.59
150	Bt	1	0	16	5	19	42	3.55	1.143	4.37	1.5	1515.17	115.03	19.89
150	Bt	2	16	41	9	20	41	3.42	1.266	4.79	1.38	2132.13	141.76	23.74
150	Bt	3	41	64	18	13	46	3.38	1.305	4.99	1.39	1685.53	103.49	20.34
150	Bt	4	64	79	14	18	40	3.52	1.437	1.85	0.55	1270.79	64.45	10.57
150	Bt	5	79	150	33	19	36	3.91	1.503	12.17	3.07	7553.54	229.89	36.68
151	A	NA	0	18	15	40	10	3.76	1.236	2.61	2.43	1968.11	155.41	31.48
151	Bt	1	18	49	11	14	42	3.81	1.331	3.8	3.52	3712.09	230.03	44.72
151	Bt	2	49	76	5	31	14	3.81	1.437	2.6	-1.25	3278.82	207.9	41.27
151	Bt	3	76	100	7	23	20	3.79	1.519	1.93	-0.92	2672.96	215.6	46.7
153	Oa	NA	0	1	90	NA	NA	NA	NA	NA	NA	NA	NA	NA
155	A	NA	0	35	50	40	32	4	0.925	10.17	6.56	1747.84	231.1	39.16
155	Bw	1	35	75	53	6	51	4.37	1.33	5.92	7.3	2671.08	213.59	27.62
155	Bw	2	75	150	65	21	37	4.22	1.55	6.52	5.9	3844.05	313.63	47
158	A	NA	0	21	9	17	47	3.99	1.337	0.4	8.44	1285.01	500.21	87.28
158	Bt	1	21	42	6	12	51	3.99	1.362	1.6	2.12	1755.37	325.94	48.29
158	Bt	2	42	54	12	11	45	3.91	1.406	0.91	0.84	931.89	140.04	22.22
158	Bt	3	54	66	8	20	27	3.94	1.382	2.09	1.72	1302.8	158.65	20.61
159	A	1	0	13	5	17	41	3.77	1.223	2.26	2.35	1303.42	119.24	29.09
159	A	2	13	29	17	12	48	3.79	1.346	0.74	1.76	1142.48	188.27	51.08
159	Bw	NA	29	74	18	8	55	4.31	1.432	2.18	23.77	1056.2	4888.77	1095.81
159	C	NA	74	108	7	18	36	4.78	1.552	1.39	26.15	743.89	7561.61	1465.46
161	A	NA	0	18	10	30	20	4.57	0.903	8.2	4.01	586.74	3232.39	232.58
161	BA	NA	18	35	6	30	20	4.26	1.313	1.9	1.95	985.45	1288.2	159.87
161	Bt	1	35	71	6	5	32	4.21	1.412	2.78	4.82	1413.17	2951.97	377.83
161	Bt	2	71	102	20	12	39	4.04	1.516	1.65	26.32	1125.43	2188.44	333.91
161	Bt	3	102	150	50	13	35	4.15	1.626	1.73	28.53	739.33	3088.83	428.89
164	Cg	1	0	14	40	10	50	4.51	1.224	1.5	1.27	391.58	298.01	39.89
164	Cg	2	14	52	37	9	57	4.53	1.268	5.58	4.9	1146.78	1112.33	141.7
164	Cg	3	52	83	30	19	49	4.57	1.332	6.51	8.66	539.22	1925.58	282.19
164	Cg	4	83	108	18	10	57	4.51	1.281	10.34	5.78	359.28	2022.29	280.36
164	Cg	5	108	150	25	11	51	4.55	1.598	5.17	2.82	1459.38	1292.5	207.76
166	A	NA	0	9	80	16	65	3.79	1.125	0.53	1.05	167.63	9.67	2.43
166	E	NA	9	80	88	9	75	4.04	1.326	1.65	3.23	732.16	53.54	9.95
166	Bw	NA	80	150	95	9	72	3.96	1.528	0.75	2.18	367.88	20.96	3.75
168	A	NA	0	13	17	24	33	3.52	1.035	4.72	1.61	1067.89	44.44	12.37
168	Bt	NA	13	36	23	20	48	3.78	1.305	3.18	1.66	1911.94	89.86	15.33
169	A	NA	0	12	8	21	38	3.33	1.199	2.31	1.26	664.03	64.83	11.18
169	Bt	1	12	47	5	17	47	3.47	1.305	5.29	1.61	4142.94	176.11	24.68
169	Bt	2	47	112	25	18	50	3.97	1.459	7.4	0.36	4476.7	337.73	52.2
169	Bt	3	112	150	16	16	61	3.97	1.6	5.51	4.01	3288.44	234.05	44.63
172	A	NA	0	26	5	29	28	3.77	1.246	4.08	2.83	2772.06	198.29	35.81
172	Bt	1	26	49	8	25	33	3.96	1.35	2.46	1.53	2804.42	127.91	23.5
172	Bt	2	49	93	12	27	34	3.89	1.442	4.9	2.15	4600.51	256.71	51.56

pedon id	horizon	suffix	top (cm)	bottom (cm)	fragvol (% volume)	clay (% weight)	sand (% weight)	pH	bulk density (g/cm^3)	C (kg/ha)	P (kg/ha)	Al (kg/ha)	Ca (kg/ha)	Mg (kg/ha)
172	Bt	3	93	150	45	25	32	3.93	1.608	2.76	2.37	3832.12	945.42	156.99
173	A	NA	0	17	14	26	47	4.62	1.191	3.27	2.18	434.2	1939.46	285.7
173	BA	NA	17	55	35	25	51	4.53	1.314	4.12	1.64	878.58	2648.89	466.48
173	Bt	1	55	80	50	13	64	4.23	1.444	1.3	1.62	343.97	1062.34	210.93
173	Bt	2	80	110	50	17	39	4.26	1.534	1.28	2.01	442.35	1709.12	368.74
173	C	NA	110	150	45	16	51	4.55	1.643	1.46	7.45	469.02	3795.52	778.08
174	A	NA	0	22	7	20	41	3.42	1.083	7.63	3.86	1667.27	482.24	55.65
174	Bt	1	22	38	13	19	42	3.98	1.292	2.54	3.45	2596.84	140.9	16.81
174	Bt	2	38	70	5	16	37	4.04	1.34	6.78	8.21	6776.69	173.75	21.54
174	Bt	3	70	109	50	12	57	4.08	1.526	1.57	4.26	2720.06	218.59	31.34
174	Bt	4	109	150	45	17	50	4.01	1.637	1.97	5.67	2344.44	355.91	48.46
175	A	NA	0	16	2	37	13	3.33	0.805	8.64	2.03	1207.92	62.28	19.43
175	BA	NA	16	42	2	32	19	3.92	1.189	7.89	4.55	4827.45	103.99	19.97
175	Bt	1	42	64	2	20	27	3.97	1.353	3.92	2.32	3885.93	100.01	16.66
175	Bt	2	64	90	2	23	28	3.84	1.455	3.3	2.13	4404.21	126.55	25.23
181	A	NA	0	8	0	32	14	4.24	1.172	2.01	1.72	247.62	914.77	103.05
181	Bt	1	8	58	0	27	15	3.87	1.362	4.87	4.39	716.22	5561.49	711.45
181	Bt	2	58	92	0	18	23	4.51	1.483	3.4	-0.53	2664.25	2195.68	316.58
181	Bt	3	92	116	37	21	21	4.64	1.546	2.13	8.8	279.5	2300.76	280.91
181	Bt	4	116	150	0	8	33	4.51	1.675	1.6	18.91	257.03	6176	1019.65
199	A	NA	0	25	5	41	24	3.78	1.09	9.11	5.09	2015.09	1090.51	143.14
199	BA	NA	25	40	2	34	24	4.24	1.262	3.72	2.37	1251.01	1171.26	192.98
199	Bt	1	40	56	10	25	28	4	1.343	2.92	0.09	1290.26	552.57	111.13
199	Bt	2	56	76	15	21	41	4.03	1.436	2.34	1.94	1293.6	477.17	103.48
400	A	NA	0	7	7	19	42	3.78	0.909	3.22	4.43	613.64	77.78	10.18
400	BA	NA	7	39	5	18	41	4.03	1.305	3.93	9.14	2919.53	207.86	30.23
400	Bt	NA	39	55	5	26	32	3.81	1.408	1.1	0.62	1514.54	63.38	14.76
400	C	NA	55	104	85	31	26	3.74	1.503	0.52	-0.55	832.29	82.46	12.67
401	A	NA	0	12	18	20	41	3.89	1.092	3.33	3.42	813.77	203.17	23.32
401	BA	NA	12	30	12	18	41	3.99	1.322	1.38	1.66	867.08	149.49	24.25
401	Bw	NA	30	46	8	14	36	3.91	1.375	1.2	-0.76	714.16	168.85	31.68
401	Bt	1	46	78	17	22	15	3.84	1.453	1.72	1.31	1545.73	483.47	112.73
401	Bt	2	78	109	40	12	34	3.89	1.532	1.59	1.35	1129.89	358.49	84.58
420	A	NA	0	14	8	32	10	4.53	1.074	4.77	0.33	674.5	2114.57	204.95
420	Bt	1	14	36	7	27	24	4.67	1.32	2.44	1.53	821.14	2555.76	307.28
420	Bt	2	36	61	21	20	42	4.46	1.399	2	-1.08	940.7	2003.55	264.47
431	A	NA	0	16	0	21	45	3.2	1.052	6.36	2.1	778.72	144.97	26.74
431	Bw	1	16	49	11	27	39	3.62	1.332	3.99	3.25	2296.29	53.03	10.11
431	Bw	2	49	90	12	20	45	3.89	1.464	3.38	-1.93	2108.36	237.33	39.31
431	Bt	NA	90	113	17	17	17	3.87	1.541	2.37	0.75	1315.04	68.69	22.98
431	BC	NA	113	150	24	16	40	3.76	1.63	3.37	-0.21	1995.77	172.21	84.05
432	A	NA	0	12	5	47	7	3.52	0.976	5.1	0.77	1462.11	31.12	17.68
432	Bt	1	12	33	5	30	10	3.61	1.267	3.51	0.82	3096.05	87.7	24.88
432	Bt	2	33	65	2	30	27	3.78	1.381	3.76	1.3	4600.14	214.65	85.01
432	Bt	3	65	91	0	38	6	3.71	1.475	2.7	0.48	4714.42	243.49	122.85
432	Bt	4	91	143	10	36	9	3.82	1.584	5.2	0.85	7672.42	444.04	191.91
434	A	NA	0	19	2	17	53	3.94	1.181	4.48	5.43	1749.73	71.72	15.7
434	Bw	1	19	40	7	21	50	3.89	1.339	1.87	1.57	1263.06	82.9	14.09
434	Bw	2	40	68	7	16	50	3.86	1.424	1.81	-1.85	1873.13	126.04	21.49
434	Bw	3	68	102	25	16	43	3.89	1.496	2.59	0.76	2362.02	156.41	55.01
434	Bw	4	102	129	30	17	31	3.87	1.556	2.98	1.04	2749.55	117.57	50.45
435	A	NA	0	28	18	23	33	3.84	1.034	10.55	12.7	2923.26	130.49	20.04
435	BA	NA	28	55	11	20	37	3.86	1.357	4.05	4.21	2281.96	141.24	25
435	Bt	NA	55	83	12	21	65	3.79	1.473	2.66	-0.4	2494.68	143.66	27.72
435	BC	NA	83	150	45	14	46	3.82	1.623	2.99	6.07	2945.8	243.26	47.25
436	A	NA	0	9	45	24	27	3.67	1.01	2.03	1.29	448.28	28.51	7.69
436	Bw	1	9	35	20	14	31	4.04	1.278	3.18	0.05	2867.64	20.64	9.54
436	Bw	2	35	70	17	12	44	4.04	1.39	3.43	1.98	2920.75	146.7	25.82
436	Bw	3	70	96	20	15	47	4.03	1.478	2.48	1.55	2187.64	113.26	16.42
436	Bw	4	96	117	50	11	57	3.99	1.553	1.11	0.83	854.64	51.5	9.44
436	BC	NA	117	140	85	5	57	4.03	1.617	0.37	0.38	279.42	17.87	3.04
437	A	1	0	30	55	8	76	4.2	1.151	3.91	11.06	1283.05	85.25	14.96
437	A	2	30	77	70	6	75	4.26	1.315	3.3	12.15	1466.2	93.2	17.01
437	BA	NA	77	112	70	3	76	3.99	1.411	3	13.22	1249.63	82.37	16.01
437	Btg	NA	112	150	45	19	32	4.26	1.625	1.95	4.33	1627.19	760.2	270.09