Digital soil mapping techniques across multiple landscape scales in South Africa

by
Trevan Flynn

Dissertation presented for the degree of
Doctor of Philosophy (Agriculture Science)
at
Stellenbosch University
Department of Soil Science, Faculty of AgriSciences

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed, and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Supervisor: Dr. Cathy Clarke
Co-supervisor: Dr. Andrei Rozanov
Co-supervisor: Dr. Willem de Clercq

December 2019

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated) that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

This dissertation includes four original papers published in peer-reviewed journals or books and one paper under review publications. The development and writing of the papers (published and unpublished) were the principal responsibility of myself and, for each of the cases where this is not the case, a declaration is included in the dissertation indicating the nature and extent of the contributions of co-authors.

Date: December 2019

Summary

Digital soil mapping has seen increasing interest due to environmental concerns and increasing food security issues. Digital soil mapping offers a quantitative approach which is cost effective as less soil observations are needed to produce large area soil maps. However, digital soil mapping has only recently been addressed in South Africa. This research aimed to produce two digital soil mapping (DSM) frameworks with the available resources in South Africa. The methodologies incorporate advanced geostatistics and/or machine learning techniques to be able to produce quantitative soil maps from the farm to catchment scale.

First, a framework that optimises both feature selection and predictive models was developed to produce farm-scale soil property maps. Four feature selection techniques and eight predictive models were evaluated on their ability to predict particle size distribution and SOC. A boosted linear feature selection produced the highest accuracy for all but one soil property. The top performing predictive models were robust linear models for gravel (ridge regression, RMSE 9.01%, R^2 0.75), sand (support vector machine, RMSE 4.69%, R^2 0.67), clay (quantile regression, RMSE 2.38%, R^2 0.52), and SOC (ridge regression, RMSE 0.19%, R^2 0.41). Random forest was the best predictive model for silt content with a recursive feature selection (RMSE 4.12%, R^2 0.53). This approach appears to be robust for farm-scale soil mapping where the number of observations is often small but high-resolution soil data is required.

Second, 24 geomorphons (landform classification) were evaluated on their association with soil classes. The geomorphon with the highest association was aggregated into a 5-unit system which was evaluated on how well the system stratified soil lightness, soil EC, SOC, effective rooting depth, depth to lithology, gravel, sand, silt, and clay. It was found that an aggregated geomorphon stratified all soil attributes except EC. Additionally, the aggregated geomorphon predicted 6 out of 9 soil properties with the greatest accuracy (RMSE) when compared to the original geomorphon (10-unit system) and a manually delineated system (5-unit system). This study shows that aggregating geomorphons can stratify the soil landscape even at the farm-scale and can be used as an initial indication of the soil spatial variability.

Third, a framework to disaggregate the Land Type Survey (LTS) through machine learning was developed. Geomorphons, together with the original LTS were overlaid to produce terrain morphological units. The polygons were disaggregated further to produce a raster map of soil depth classes through a disaggregation algorithm known as DSMART. The first most probable class raster achieved an accuracy of 68% and for the two most probable class rasters, an accuracy of 91% was achieved. The two-step approach proved necessary for producing a farm-scale soil map.

Forth, a study aimed to compare 10 algorithms, implemented through a modified DSMART model, in their ability to disaggregate two polygons into soil associations in two environmentally contrasting locations (Cathedral Peak, KwaZulu-Natal Province and Ntabelanga, Eastern Cape Province). At Cathedral Peak (high relief with clear toposequences), nearest shrunken centroid was the top performing algorithm with a kappa of 0.42 and an average uncertainty of 0.22. At Ntabelanga (low relief with strong geological control), the results were unsatisfactory. However, a regularised multinomial regression was the top performing algorithm, achieving a kappa of 0.17 and an average uncertainty of 0.84. The results of this study highlight the versatility of a technique to disaggregate South Africa's national resource inventory.

Disaggregation was then used to simultaneously disaggregate 20 land types in the Mvoti catchment covering 317 km^2 in KwaZulu Natal province. First, the optimal geomorphon was chosen through a spatially resampled Cramer's V test to determine the association between the soil legacy polygons and the geomorphon units. Second, feature selection algorithms were embedded into DSMART. Third, the feature selection techniques were compared using 25, 50, 100, and 200 resamples per polygon. The results indicate that the Cramer's V test is a rapid method to determine the optimal input map. Feature selection algorithms achieved the same accuracy as using all covariates but had greater computational efficiency. It is recommended that 10 to 20 times the amount of soil classes be used for the number of resamples per polygon.

This dissertation is dedicated to my parents:

Anne Bennett and Michael Flynn

Biographical sketch

I, Trevan Flynn, graduated from Stellenbosch University (cum laude) on 8th December 2015 with a BScAgric (Soil Science and Horticulture). In 2017, I went on to do a MScAgric (Soil Science) where it was upgraded to a PhD (Soil Science) on 14th November 2018 by the Executive Committee of Senate of the Faculty of Agriculture. During my MSc, I also developed and lectured an introductory digital soil mapping course for fourth year soil science students at Stellenbosch University. Since the beginning of my PhD, I have also been working with BetterWorld Energy and the Zambian Institute of Agriculture. My activities include developing a protocol and models to supply smallholder farmers with soil analysis and where I lecture introductory soil science, crop science and agro-climatoloty, respectively. My research focuses on using existing and freely available data to increase the accessibility and accuracy of soil spatial information.

Acknowledgements

I wish to express my sincere gratitude and appreciation to the following persons and institutions:

- Dr Cathy Clarke for encouraged me to continue to study, supervising both my MSc and PhD, as well as supporting my research ideas. My upgrade to PhD could not have been done without your help which assisted me and my career substantially.

- Dr Andrei Rozanov for supervising my research, helping me develop new research ideas, and helping in writing all papers published. Your help in my life outside academics was truly great and I appreciate all your everything.

- Dr Willem de Clercq who co-supervised my MSc and who helped me understand the spatial variability of soils.

- Dr Lisel Wise who also helped me outside academics and who trusted my skill set to help in your PhD thesis. I hope we can write more papers together.

- Dr Benjamin Warr who hosted me in Zambia, helped me understand spatial statistics, and co-authored two papers. Thank you for letting me be a part of your projects.

- To Michael Flynn and Karen Angus who helped edit my first publication.

- Finally, to my parents who supported me through this whole process. The support was truly amazing, even though I was half way across the world.

Preface

This dissertation is presented as a compilation of 7 chapters where Chapters 2 – 6 are presented in research paper format and have been either submitted, under-review, or have been published in peer-reviewed journals. A declaration is placed in each chapter where co-authors contributions have been stated.

The first chapter gives a short introduction to digital soil mapping, the problem statement, and the aim of the research. Chapter 2 introduces a digital soil mapping framework that optimizes covariate selection and predictive models in the Swartland, Western Cape. This chapter has been published by the *South African Journal of Plant and Soil* and supplementary material was presented at the 21st World Soil Congress in Rio de Janeiro, Brazil. On the same location, chapter 3 uses landform elements to stratify the soil-landscape into meaningful patterns. Chapter 3 has been accepted for publication in the Journal *Catena*. Chapter 4 introduces a digital soil mapping technique to disaggregate South Africa's land types into a farm-scale soil depth map. This chapter has been published in the journal *Geoderma*. Chapter 5 introduces further upgrades to the framework developed in Chapter 4 which includes the incorporation of additional machine learning algorithms and optimization features. This was conducted in two environmentally contrasting land types at Cathedral Peak, KwaZulu Natal and Ntabelanga in the Eastern Cape. This chapter has been published in the journal *Geoderma*. Chapter 6 simultaneously disaggregates multiple land types in KwaZulu Natal Midlands. This chapter addresses drawbacks identified in Chapter 4 and 5, namely, computational efficiency and final accuracy. This chapter is being submitted to the journal *Geoderma*. The final chapter gives a general conclusion and further work that is needed.

As the thesis is presented in paper format, there will be some natural repetition; however, the papers have been modified to reduce this repetition.

Table of Contents

List of Figures

List of Tables

Symbols and abbreviations

ALOS -2:	Advanced land observation satellite
BI:	Brightness index
C5:	C5 decision trees
cLHS:	Conditioned Latin hypercube sampling
CI:	Colouration index
DEM:	Digital elevation model
DSM:	Digital soil mapping
DSMART:	Disaggregating and harmonizing soil map units through resample classification trees
FSAs:	Feature selection algorithms
KNN:	K-nearest neighbour
LOOCV:	Leave-one-out cross validation
LFEs:	Landform elements (2 or 3-dimensional system)
LDA:	Linear discriminatory analysis
LTS:	South Africa Land Type Survey
LTS-GM:	Vectorised geomorphon map overlaid with the Land Type Survey
GLS:	Generalised least squares
GRASS:	Geographic Resources Analysis Support System
NSC:	Nearest shrunken centroid
MLP:	Multiple layer perceptrons
NIR:	Near infrared
REI:	Relative efficiency index for feature selection techniques
REML:	Residual maximum likelihood analysis
RF:	Random Forest
RR:	Ridge regression
RMSE:	Root mean squared error
SAGA:	System for Automated Geoscientific Analysis
SGB:	Stochastic gradient boosting
SI:	Saturation index
SVL:	Linear kernel support vector machines
SVR:	Radial kernel support vector machines
TMU:	Terrain morphological units (2-dimensional system)

List of publications

List of publications included in the thesis were:

Flynn, T., Clarke, C., de Clercq, W., Rozanov, A. 2019. "High Resolution Digital Soil Mapping on a Complex Landscape in the Sandspruit catchment, Western Cape, South Africa," Proceedings of the 21st World Congress of Soil Science, Aug. 12-17, 2018, pp. 234.

Flynn, T., de Clercq, W., Rozanov, A., Clarke, C. 2019. High-resolution digital soil mapping of multiple soil properties: an alternative to the traditional field survey? *South African J. Plant Soil*. https://doi.org/https://doi.org/10.1080/02571862.2019.1570566.

Flynn, T., Rozanov, A., de Clercq, W., Warr, B. & Clarke, C. 2019. Semi-automatic disaggregation of a national resource inventory into a farm-scale soil depth class map. *Geoderma* 337, 1136-1145.

Flynn, T., van Zijl, G., van Tol, J., Botha, C., Rozanov, A., Warr, B. & Clarke, C. 2019. Comparing algorithms to disaggregate complex soil polygons in contrasting environments. *Geoderma* 352, 171 -180.

(Accepted for publication) Flynn, T., de Clercq, W., Rozanov, A., Clarke, C., Farm-scale soil patterns derived from automated terrain classification. *Catena*.

(Under review) Flynn, T., Rozanov, A., Clarke, C., Input map and feature selection for soil legacy data. *Geoderma*.

As the lead author of all these manuscripts, the majority of the content contribution was my own. The co-authors contributed with the writing aspect, critical review and understanding the soil resource available. There is one paper accepted for publication in the journal *Catena* and one paper under review in the journal *Geoderm*a. Each article can be found in Appendix A through C. The contribution of all co-authors was extremely appreciated.

Chapter 1 Introduction

1.1 Introduction

As of 2015, there are 2.6 million smallholder farming and 350,000 commercial farming households in South Africa (Pienaar and Traub, 2015). The majority of smallholder farmers come from former "homelands" in and have been previously disadvantaged through past socio-economic-political policies (Kirsten and van Zyl, 1998). Due to past policies, there is a major lack of land use management knowledge and very limited access to resources such as soil information which persists to this day. This has contributed towards low yields and land degradation creating food and economic insecurity in rural areas.

In drought prone South Africa, these issues are perpetuated during the El Niño weather pattern effecting farmers as well as urban locations such as Cape Town in the Western Cape Province, South Africa. For example, harsh water restrictions were placed in the Western Cape during the 2015-2017 drought and the reservoirs that supply Cape Town with water, were depleted. Therefore, spatial soil information is crucial to understanding such things as ground water recharge for hydrological models as soils redistribute precipitation on the soil surface as well as in the subsoil (Park et al., 2001). Spatial soil information is often the most limiting factor in hydrological modelling (Wahren et al., 2016) and highly detailed soil information is required for rainfall-runoff models such as the J2000 model (Krause, 2001).

The solution lies in a detailed soil survey and/or digital soil mapping (DSM) techniques. However, conventional soil maps have a variety of limitations, for example they have discrete boundaries, represented in a polygon format (Zhu, 1997). Burrough (1989), considers this a major loss of soil information as conventional maps disregard polygon homogeneity, spatial variability, and soil map errors (measurement and spatial). Furthermore, conventional soil surveys are time consuming, expensive, and impractical in remote regions (Bui and Moran, 2003).

Digital soil mapping on the other hand, is the quantitative prediction of soil classes and continuous soil properties achieved by utilising recent advancements in highly detailed ancillary data, computational power, Geographic Information Systems (GIS) software, as well as geostatistical and machine learning algorithms (Scull et al., 2003). In other words, DSM is the computer assisted production of soil maps where the accuracy of the maps can be assessed.

Digital soil mapping has been developing since the 1970s, but can be dated back to 1911 (Webster, 1994) in the agronomical study by Mercer and Hall (1911). The study examined the crop yield variation between plots and explored geostatistical features such as autocorrelation. In South Africa, Kantey and Williams (1962) used computer-based models to map soils and determine viable road projects as well as to evaluate the financial cost of such maps. According to Webster (1994), they were also the first to validate the produced maps. Unfortunately, the techniques used by Kantey and Williams were not further pursued in South Africa until(van den Bergh and Weepener, 2008) who mapped zones of high agriculture potential in KwaZulu Natal, Province.

Since the 1970s, DSM has developed into a common framework that has produced satisfactory accuracies when mapping soil resources for large areas. The framework consists of correlating soil attributes with environmental factors and predicting the soils over the area of interest. Such research/projects include GlobalSoilMap (Arrouays et al., 2014) and SoilGrids (Hengl et al., 2017, 2014). However, there has been comparatively little DSM research in South Africa where the DSM framework needs to be adapted to address local needs (van Zijl, 2019). Questions still remain if DSM can help solve these needs as very little research into DSM has been conducted in South Africa. A major concern is the lack of DSM research and skillsets, which are vital as to address these issues (Paterson et al., 2015)

Internationally, farm-scale DSM has only moderately been addressed due to a lack of financial resources and expertise (Xu et al. 2018). The farm-scale DSM research that has been conducted, predominantly has focused on precision agriculture for large commercial farms. For example, Iticha and Takele (2019) mapped soil texture, pH, acidity, and nutrients for a 40 km^2 area of agriculture land in Ethiopia using a geostatistical approach. Triantafilis et al. (2000), mapped cation exchange capacity (CEC, *cmol (+)/kg*) using proximal and remote sensing imagery for a 1,500 km^2 cotton farm. However, are these studies relevant in South Africa which experience a high soil spatial variability within short distances and where the average farm size is 264 ha (Wet, 2018)?

Nevertheless, highly detailed soil information at the farm-scale is important to reduce the environmental impact of agriculture such as avoiding marginal land use, over application of fertilizer, and over irrigating (especially in drought prone regions). Digital soil mapping can potentially supply

this information through freely available resources such as satellites and software. Therefore, research into DSM techniques should be further investigated.

The predominant issue facing farm-scale DSM is the small number of soil observations on farms or there are no observations at all in the case of previously disadvantaged farmers. Small data sets are problematic because; i) it is difficult to find soil-environmental relationships, ii) machine learning algorithms are data driven, and iii) unimportant variables and/or collinear variables can add bias to the model (Kuhn and Johnson, 2013). This affects the interpretability and accuracy of the produced maps making them misleading (Mason and Perreault 1991).

In a country such as South Africa, where resources for collecting soil information are limited, it makes economic sense to maximise the use of existing datasets. Paterson et al. (2015), states that priorities of soil information in South Africa include "*interdisciplinary collaboration; expansion of the current national soil database with advanced data acquisition, manipulation, interpretation and countrywide dissemination facilities*". There are two digital soil databases that cover all of South Africa. These soil databases include the Land Type Survey (LTS) at a 1:250,000 scale (Land Type Survey Staff, 1972 - 2006) and the Soil and Terrain Database for South Africa (SOTER) at a 1:1 million scale (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012). However, the SOTER database used the LTS in its creation, thus will not be the focus of this study.

The LTS is a country wide soil, terrain, and macroclimate map which defines the agricultural potential (in terms of soil, climate, and terrain) of South Africa (Land Type Survey Staff, 1972 - 2006). The LTS took many years to create and multiple soil scientists were involved in its creation. First, landscape units of uniform drainage pattern, profile type, percentage level land and local relief were delineated for areas visible at 1:50 000 scale. In the LTS, such land scape units are known as terrain types. These terrain types were further divided into terrain morphological units (TMUs) representing the 5-unit landscape model of 1-crest, 2-scarp, 3-midslope, 4-footslope and 5-valley bottom. Most often terrain types include a repeating pattern of TMUs, for example crest-valley sequences that repeat themselves. Soil information was collected on representative TMUs and recorded in an inventory representing the percentage area that a soil type covers a TMU. Final land type units were constructed by overlaying a macroclimate map so that units represented areas of uniform terrain, repeating soil pattern and macroclimate (Land Type Survey Staff, 1972 - 2006).

Due to the country wide coverage and amount of information contained, van Zijl et al. (2013), states that DSM in South Africa should focus on disaggregating the LTS. Disaggregation involves the spatial placement of individual soil components from soil map polygons (McBratney, 1998). Research into disaggregation techniques has been limited, however, interest has recently grown because it does not require an additional soil survey to produce soil maps. Some recent studies include Møller et al.,(2019), who disaggregate two different national soil maps in Denmark, Chaney et al. (2016) who disaggregated the contiguous United States using the Soil Survey Geographic (SSURGO) database, and Zeraatpisheh et al. (2019) who disaggregated soil legacy polygons in Chaharmahal-Va-Bakhtiari Province in central Iran. Disaggregation techniques are especially relevant to the LTS where there are only a few georeferenced soil observations per land type while some land types do not have any soil observations. Additionally, disaggregation does not reject the knowledge used in creating the LTS, rather, disaggregation uses the existing knowledge to create more detailed and reproducible soil maps.

Given that terrain mapping is the backbone of the LTS, any disaggregation technique requires automated TMU delineation to standardise the disaggregation process. Standardisation will help make the disaggregation of the LTS more reproduceable and quantitative. Additionally, standardisation would help with data acquisition and expand the LTS detail on different landscapes and scales. This can be achieved using digital elevation models (DEMs). Digital elevation models can be derived from ground surveys, light detection and ranging (LiDAR), or satellite images (Nelson et al., 2009). A lot of satellite imagery is freely available down to a resolution of 12.5 m making them the most cost-effective method. Therefore, the use of satellite images can help further reduce the cost of the final map.

There are many possible ways to predict TMUs that are represented in the LTS. One of the most recent advancements is known as geomorphons developed by Jasiewicz and Stepinski (2013). Geomorphons are a pattern recognition algorithm that classifies the 10 most common landform elements from 498 possible features. Landform elements are topographic features such as summit, slopes, and valleys. The dominant advantage of the geomorphon algorithm is that it is computationally efficient making it possible to map whole countries (Jasiewicz and Stepinski, 2013).

Additionally, geomorphons are robust in terms of scale and can be aggregated to represent landscape position (Libohova et al., 2016).

Given the reliance of the LTS on terrain delineation, the geomorphon approach is appealing, however the 10-unit geomorphon model would need to be converted to a 5-unit model to see how well it stratifies the soil-landscape compared to the LTS data. This is important because if geomorphons relate to soil properties they could potentially be used as an indication of the soil distribution. It is also useful to establish how the 5-unit geomorphon model compares to the "mental model" of the surveyor who constructed the LTS in a specific region. If geomorphons do, the existing knowledge of the spatial soil distribution could be extracted.

Disaggregation techniques often produce less satisfactory results in terms of accuracy when compared to using point observations in machine learning algorithms. This is because, in general, there are no georeferenced soil observations used in the predictive model during disaggregation. Disaggregation techniques either use expert knowledge to create soil-environmental rules or use algorithms which rely on soil class probabilities in the soil legacy legends. Therefore, it is more difficult and/or time consuming to find soil-environmental relationships. This is a necessary yet detrimental feature of disaggregation techniques. For example, van Zijl et al. (2013) found a 32% decrease in accuracy when disaggregating two land types without soil observations. However, this helps map remote locations and reduces the cost needed to produce a soil map.

Modern disaggregation techniques such as DSMART ("Disaggregating and Harmonizing Soil Map Units Through Resampled Classification Trees") extracts the expert knowledge of the soil surveyor by using machine learning algorithms to predict soils. Developed by Odgers et al. (2014), DSMART randomly samples polygons and assigns soil classes based on the soil legacy legend. The DSMART algorithm then trains a decision tree (type of machine learning algorithm) and predicts the model over the area of interest. This is done in an iterative process to create a user defined number of realisations of the soil distribution. Either the realisation with the highest accuracy can be used for the final map (deterministic approach) or a probability series of soil class rasters can be created (stochastic approach) by counting the number of times each pixel is classified as a specific soil class. The stochastic approach could have benefits when disaggregating places with low relief where the soil distribution appears random. However, the former approach is easier to interpret.

Most disaggregation techniques, such as DSMART are implemented with a coarse resolution (> 20 m) on large geographic regions such as studies by Holmes et al. (2015) and Chaney et al. (2016). However, this offers little use to farmers as the resolution of the final map is too coarse for smallholder farmers and precision agriculture (McBratney et al., 2000). Locally and internationally, the use of national soil databases to produce farm-scale soil maps has yet to be researched. As this research will show, disaggregation techniques produce better results when there are clear toposequence and soil-environmental patterns. This aspect must be considered when disaggregating the interior of central South Africa where there is generally low relief and depending on the scale needed, parent material might be the main driver of the soil spatial variability. However, geological information is not displayed in either the LTS or SOTER.

Despite its many benefits the DSMART approach has many parameters to optimise such as the number of resamples per polygon, number of realisations, and if other algorithms will perform better. Therefore, a modified DSMART method that is capable of optimising algorithm selection and model parameters is warranted to produce soil maps on different landscapes, purposes, and scales. Although DSMART has the benefit of unlocking expert knowledge contained in the legacy maps, the computational, the computational intensity also limits its accessibility. Therefore, a method that helps reduce this computational intensity is also essential, especially when mapping large regions at a fine resolution.

1.2 Background

1.2.1 Common DSM framework

In DSM, it is assumed soils or their characteristics, have a correlation with known environmental attributes. A predictive model learns these relationships and then predicts soils at unobserved locations. Thus, soil attributes are a function of environmental data known as covariates which are usually represented in a raster format (pixels) (McBratney et al., 2003). The raster format helps display the continuous nature of soils and the factors that determine their spatial variability in an easily understandable format (Zhu, 1997).

The mathematical representation of this is shown in Equation 1.1, where S is the soil attribute of interest, f is usually a deterministic machine learning algorithm (known relationships), Q are

covariates, and ε is the random error that is usually accounted for using a stochastic geostatistical models (unknown relationships). The machine learning algorithms can either be multilinear or nonlinear models.

$$S = f(Q) + \varepsilon \tag{1.1}$$

The function f, finds relationships with Q and predicts soils using relationships learnt. The predictions of soil properties is made through geostatistics, machine learning or a hybrid between the two (McBratney et al., 2003). Each method has its own advantage and the optimal method will depend on soil attribute being predicted (Forkuor et al., 2017), spatial resolution required (Hengl, 2006), geographic location (Brungard et al., 2015) as well as scale (Möller and Volk, 2015).

The advantage of a purely geostatistical approach such as kriging, is that it is the best unbiased linear estimate of soil properties (Matheron, 1963). In other words, the soil samples taken at a specific location will not change value as might happen in a linear regression. Additionally, geostatistics considers spatial autocorrelation which should always be checked when analysing spatial data. However, the data must have a normal distribution and have a homogenous variance. Kriging achieves the best accuracy when the soil samples are systematically aligned (gridded) as done in traditional soil surveys.

Some machine learning algorithms on the other hand, do not assume a normal distribution or homogenous variance such as tree-based models. Additionally, machine learning can be very efficient at finding soil-environmental relationships that would otherwise go unobserved by a soil surveyor (Kuhn and Johnson, 2013). However, machine learning requires a large amount of soil data and at the farm-scale, soil data with a very low error. For example, soil measurements/classification, georeferenced soil profiles, and environmental data all need to be of high quality (low errors). Machine learning works best with a randomised soil sample design to prevent bias.

The covariates Q are related to Jenny (1941)'s five soil forming factors of climate, vegetation, topography, parent material, and age. McBratney et al. (2003), later added easily measured soil attributes and neighbourhood to the covariates as well as the random error term in what's known as the *scorpan-SSPFe (scorpan* framework*; "soil spatial prediction function with spatially autocorrelated*

errors"). The mathematical representation of the *scorpan* framework is $S = f(s,c,o,r,p,a,n) + \varepsilon$. Where, *S* is the soil attribute of interest, *s* is easily measured soil properties, *c* is climate factors, *o* is organisms (e.g., vegetation), *r* is relief, *p* is parent material and *n* is neighbourhood (i.e., spatial position). The inclusion of additional soil properties and location indicates that soils can be predicted through other (easily measured) soil properties (McBratney and Webster 1983), geographic location (Lagacherie 1992), and location relative to important features such as distance to stream (Bui and Moran 2000). The *scorpan* framework does not indicate that these factors are influencing the formation of soils in the area, only that these covariates correlate to the spatial soil distribution. This is a necessary distinction as DSM is not a pedogenic model, it is a spatial predictive model.

The random error term ε (residuals of f), should have a normal distribution to conform with the assumptions of general geostatistics. This term is usually accounted for using a variogram to determine the spatial autocorrelation and then kriged (Matheron, 1963). The kriged residuals are then added back to the deterministic model (hybrid model) to account for any spatial autocorrelation. This method is known as regression kriging (Odeh et al., 1994). The random error is usually accounted for soil continuous properties, not soil classes, and the implementation should be interpreted with caution. The caution arises from the variogram model used, the number of soil observations, the sample design, the number of observations, and what the need is for interpolation (Hengl and Evans, 2009). A review of geostatistics can be found in Webster and Oliver (2007), however, due to the complexity of geostatistics, a review is beyond the scope of this thesis.

1.2.2 Soil observations

There are two types of soil data that can be used to find soil correlations with covariates. Georeferenced soil observations (point) is the most commonly used soil data. Soil point data is obtained from an existing soil survey or a new survey can be conducted. Advantages of a new survey are that they can be designed to be random and cover the spatial and environmental variability of an area. Therefore, the sampling design can be optimised for the soil attribute of interest, scale, and geographic location. There are many sampling strategies that can be implemented including simple random sampling, geostatistical approaches (Brus and de Gruijter, 1997; Vašát et al., 2010), spatial coverage sampling (Royle and Nychka, 1998), stratified random sampling (Lacoste et al., 2014), define a reference area (van Zijl et al., 2019), and conditioned Latin Hypercube sampling (Minasny and McBratney, 2006).

There have been many studies that use soil point observations to predict soil attributes. These attributes include soil types (Brungard et al., 2015; Jafari et al., 2012; Jafari et al., 2014; da Silva et al.,2014; Stum, 2010; Zhu, 1997), soil organic matter (Chabala et al., 2017; Hoffmann et al., 2014; Mondal et al., 2017; Mora-Vallejo et al., 2008; Ray et al.,2004; Riggers et al., 2019), total N (Morellos et al., 2016; Were et al., 2015), P content (Mckenzie and Ryan, 1999), particle size distribution (Forkuor et al., 2017; Román Dobarco et al., 2017; Zhang et al., 2013), soil depth (Boer et al., 1996; Hengl et al., 2015; Leenaars 2018; Menezes et al., 2014), cation exchange capacity (Forkuor et al., 2017; Hengl et al., 2017), and more. These predictions have been made through a number of geostatistical and machine learning approaches.

Geostatistics can be accounted for using point observations. This is because the distance between each observation can be calculated with precision. It is common to check the residuals for any spatial autocorrelation if using a hybrid method. If the deterministic model's residuals show spatial autocorrelation than this must be accounted for. If not, the model will have bias, it will be difficult to interpret, and the accuracy will decrease (Hengl and Evans, 2009).

1.2.3 Soil legacy data

Soil legacy maps are another type of soil data which can be used in the DSM framework. This data is in the form of polygons which were drawn at a certain scale and purpose. These surveys are found in almost every country and to this day, serve as the main source of soil data (Arrouays et al., 2017). However, polygons can consists of multiple soil classes and therefore, must be disaggregated into individual soil components to achieve a highly detailed map (Odgers et al., 2014).

Soil attributes which have been predicted through disaggregation techniques predominantly focus on soil classes such as fluvial facets (Bui and Moran, 2001), soil types (Häring et al., 2012; Holmes et al., 2015; Møller et al., 2019; Nauman and Thompson, 2014; Odgers et al., 2014; Vincent al., 2018), soil associations (Van Zijl et al., 2013), and as this study will show, soil depth classes. The prediction of soil classes is due to the nature of soil legacy maps which heavily rely on soil types. However, soil properties, such as soil pH have also been predicted from the disaggregated soil class probabilities (Odgers et al., 2015). These predictions have been made from environmentally clustering (Bui and Moran, 2001; Zeraatpisheh et al., 2019), expert knowledge (van Zijl et al., 2013), as this study will

show multinomial logistics regression and many other algorithms, decision trees (Bui and Moran, 2001; Häring et al., 2012; Odgers et al., 2014) and/or complex disaggregation algorithms (Chaney et al., 2016; Møller et al., 2019; Nauman and Thompson, 2014).

Geostatistics is not commonly used in disaggregation techniques. This is for two important reasons. First, soil classes are the primary attribute being predicted which there are no residuals to account for (random error term). Second, the geographic distance between each class is difficult to calculate as within a polygon, there is a large room for error when allocating a soil class. However, Kerry et al. (2012) used area to point kriging to disaggregate legacy data into soil organic carbon content. Alternatively, coordinates can be used as a covariate to account for neighbourhood in the *scorpan* factors or spatial distance buffers between soil classes could be used as in the RFsp (Random Forest for spatial data) framework (Hengl et al., 2018).

These DSM techniques are necessary in order to improve the quality of soil information in South Africa. The following chapters explore the use of existing soil information sources in South Africa and offer advancements in the DSM approach to start addressing land use management and environmental needs.

1.3 Aims and objectives

The aim of this study was to develop two DSM frameworks that use freely available ancillary data and software to increase accessibility of soil spatial information to farmers and environmentalists.

One framework (framework 1) was to develop DSM techniques to produce farm-scale soil maps and meaningful soil-landscape patterns through point observations. The main objectives were:

Objective 1 (Chapter 2): Develop a DSM framework that optimises covariate selection and predictive models simultaneously to produce soil maps of multiple soil properties. This method differs from the common DSM framework as it specifically places covariate selection into the methodology.

Objective 2 (Chapter 3): Compare how soil properties and classes are stratified by TMUs delineated from expert knowledge, a 10-unit, and 5-unit geomorphon model. This was done to determine if geomorphons are a good indication of spatial soil properties.

The second framework (framework 2) was to develop a methodology to disaggregate the LTS from the farm to catchment scale. This will help improve the soil information in South Africa and its accessibility. The main objectives were:

Objective 3 (Chapter 4): Apply the DSMART technique to disaggregate the LTS into a meaningful farm-scale soil depth map comparing an aggregated geormorphon, a manually delineated LFE system, and using the original LTS polygons as input maps.

Objective 4 (Chapter 5): Compare the capacity of DSMART to disaggregate contrasting pedological environments (in terms of relief and parent material) and compare numerous algorithms for the DSMART implementation in each region.

Objective 5 (Chapter 6): Optimisation of the DSMART technique to streamline selection of input maps, feature selection and resampling rates to reduce the computational efficiency and quantitatively select covariates. This can increase the accessibility, interpretability, and accuracy of the approach.

These objectives are dealt with in Chapters 2-6 in the order outlined above. These chapters represent the core research of this thesis. Chapters 2-6 are based on manuscripts that have either been published or are under review. Chapter 7 highlights the major conclusions of the study and suggests further work.

1.4 References

Arrouays, D., Leenaars, J.G.B., Richer-de-forges, A.C., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., Heuvelink, G., Batjes, N., Carvalho, E., Hartemink, A., Hewitt, A., Hong, S., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A., Mcbratney, A., Mckenzie, N., Vasquez, G.M., Leatitia, V., Minasny, B., Montanarella, L., Odeh, I., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Solbovoy, V., Thompson, J., Smith, S., Sulaeman, Y., Vintila, R., Viscarra, R., Wilson, P., Zhang, G., Swerts, M., Oorts, K., Karklins, A., Feng, L., Ibelles, A.R., Levin, A., Laktionova, T., Dell, M., Suvannang, N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallett, S., Keay, C., Farewell, T., Lilja, H., Juilleret, J., Marx, S., Takata, Y., Kazuyuki, Y., Mansuy, N., Panagos, P., Liedekerke, M. Van, Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Kacem, S., Moussadek, R., Badraoui, M., Da, M., Paterson, G., Gonçalves, C., Theocharopoulos, S., Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozák, J., Boruvka, L., Dobos, E., Taboada, M., Moretti, L., Rodriguez, D., 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14, 1–19. https://doi.org/10.1016/j.grj.2017.06.001

Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., McBratney, A.B. (Eds.), 2014. GlobalSoilMap: Basis of the global spatial soil information system, in: 1st GlobalSoilMap Conference. CRC Press, Orleans, France.

Boer, M., Barrio, G. Del, Puigdefiibregas, J., 1996. Mapping soil depth classes in dry Mediterranean areas using terrain attributes derived from a digital elevation model. Geoderma 72, 99–118.

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

Brus, D.J., de Gruiter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion) 1, 45–49.

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

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

Bui, E.N., Moran, C.J., 2000. Regional-scale investigation of the spatial distribution and origin of soluble salts in central north Queensland. Hydrol. Process. 14, 237–250.

Burrough, P.A., 1989. Fuzzy mathematical methods for soil survey and land evaluation. J. Soil Sci. 40, 447–492. https://doi.org/https://doi.org/10.1111/j.1365-2389.1989.tb01290.x

Chabala, L.M., Mulolwa, A., Lungu, O., 2017. Application of Ordinary Kriging in Mapping Soil Organic Carbon in Zambia. Pedosphere 27, 338–343. https://doi.org/10.1016/S1002-0160(17)60321-7

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025

da Silva, A.F., Pereira, M.J., Cameiro, J.D., Zimback, C.R.L., Landim, P.M.B., Soares, A., 2014. A new approach to soil classification mapping based on the spatial distribution of soil properties. Geoderma.

FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012. Harmonized World Soil Database (version 1.2). Rome, Italy.

Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M., 2017. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. PLoS One 12. https://doi.org/10.1371/journal.pone.0170478

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma 185–186, 37–47. https://doi.org/10.1016/j.geoderma.2012.04.001

Hengl, T., 2006. Finding the right pixel size. Comput. Geosci. 32, 1283–1298. https://doi.org/10.1016/j.cageo.2005.11.008

Hengl, T., De Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE. https://doi.org/10.1371/journal.pone.0169748

Hengl, T., De Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km - Global soil information based on automated mapping. PLoS One 9. https://doi.org/10.1371/journal.pone.0105992

Hengl, T., Evans, I.S., 2009. Mathematical and digital models of the land surface. Dev. Soil Sci. 33, 31–63. https://doi.org/10.1016/S0166-2481(08)00002-0

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., De Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One 10, 1–26. https://doi.org/10.1371/journal.pone.0125814

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, 2–49. https://doi.org/10.7717/peerj.5518

Hoffmann, U., Hoffmann, T., Jurasinski, G., Glatzel, S., Kuhn, N.J., 2014. Assessing the spatial variability of soil organic carbon stocks in an alpine setting (Grindelwald, Swiss Alps). Geoderma 232–234, 270–283. https://doi.org/10.1016/j.geoderma.2014.04.038

Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. CSIRO 53, 865–880.

Iticha, B., Takele, C., 2019. Digital soil mapping for site-specific management of soils. Geoderma 351, 85–91. https://doi.org/10.1016/j.geoderma.2019.05.026

Jafari, A., Finke, P.A., Vande Wauw, J., Ayoubi, S., Khademi, H., 2012. Spatial prediction of USDA- great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types. Eur. J. Soil Sci. 63, 284–298. https://doi.org/10.1111/j.1365-2389.2012.01425.x

Jafari, A., Khademi, H., Finke, P.A., Van de Wauw, J., Ayoubi, S., 2014. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. Geoderma 232–234, 148–163. https://doi.org/10.1016/j.geoderma.2014.04.029

Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons - a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156. https://doi.org/10.1016/j.geomorph.2012.11.005

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw- Hill, NY. https://doi.org/10.2307/211491

Kantey, B.A., Williams, A.B.A., 1962. The use of soil engineering maps for road projects. Transvaal South African Inst. Civ. Eng. 4, 149–159.

Kerry, R., Goovaerts, P., Rawlins, B.G., Marchant, B.P., 2012. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. Geoderma 170, 347–358. https://doi.org/10.1016/j.geoderma.2011.10.007

Kirsten, J., van Zyl, J., 1998. Defining small-scale farmers in the South African context. Agrekon 37, 551–562. https://doi.org/10.1080/03031853.1998.9523530

Krause, P., 2001. Das hydrologische Modellsystem J2000: Beschreibung und Anwendung in großen Flußgebieten. Umwelt/Environment 29.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296–311. https://doi.org/10.1016/j.geoderma.2013.07.002

Lagacherie, P., 1992. Formalisation des lois de distribution des sols pour automiser la cartographie pedologique a partir d'un secteur pris comme reference. Universite Montpellier II, France.

Land Type Survey Staff, 1972-2006. Land Types of South Africa on 1:250 000 scale. Pretoria, South Africa.

Leenaars, J.G.B., Claessens, L., Heuvelink, G.B.M., Hengl, T., Ruiperez, M., Bussel, L.G.J. Van, Guilpart, N., Yang, H., Cassman, K.G., 2018. Mapping rootable depth and root zone plant-available water holding capacity of

the soil of sub-Saharan Africa. Geoderma 324, 18–36. https://doi.org/10.1016/j.geoderma.2018.02.046

Libohova, Z., Winzeler, H.E., Lee, B., Schoeneberger, P.J., Datta, J., Owens, P.R., 2016. Geomorphons: Landform and property predictions in a glacial moraine in Indiana landscapes. Catena 142, 66–76. https://doi.org/10.1016/j.catena.2016.01.002

Mason, C.H., Perreault, W.D., 1991. Collinearity, Power, and Interpretation of Multiple Regression Analysis. J. Mark. Res. 28, 268–280.

Matheron, G., 1963. Principles of geostatistics. Soc. Econ. Geol. 58, 1246–1266. https://doi.org/https://doi.org/10.2113/gsecongeo.58.8.1246

McBratney, A.B., 1998. Some Considerations on methods for spatially aggregating and disaggregating soil informtion, in: Soil Water Quality at Different Scales. Developments in Plant and Soil Science. pp. 51–62.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327. https://doi.org/10.1016/S0016-7061(00)00043-4

McBratney, A.B., Santos, M.L.M., Minasnya, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

McBratney, A.B., Webster, R., 1983. Optimal interpolation and isarithmic mapping of soil properties: V. Co-regionalization and multiple sampling strategy. Eur. J. Soil Sci. Soil Sci. 34, 137–162.

Mckenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94.

Menezes, M.D. de, Silva, S.H.G., Mello, C.R. de, Owens, P.R., Curi, N., 2014. Solum depth spatial prediction comparing conventional with knowledge-based digital soil mapping approaches. Sci. Agric. 71, 316–323. https://doi.org/10.1590/0103-9016-2013-0416

Mercer, W.B., Hall, A.D., 1911. The Experimental Error of Field Trials. J. Agric. Sci. 4, 107–132. https://doi.org/10.1017/S002185960000160X

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

Møller, A.B., Malone, B., Odgers, N.P., Beucher, A., Vangsø, B., Humlekrog, M., Minasny, B., 2019. Improved disaggregation of conventional soil maps. Geoderma 341, 148–160. https://doi.org/10.1016/j.geoderma.2019.01.038

Möller, M., Volk, M., 2015. Effective map scales for soil transport processes and related process domains - Statistical and spatial characterization of their scale-specific inaccuracies. Geoderma 247–248, 151–160. https://doi.org/10.1016/j.geoderma.2015.02.003

Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., Mukhopadhyay, A., 2017. Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data. Egypt. J. Remote Sens. Sp. Sci. 20. https://doi.org/10.1016/j.ejrs.2016.06.004

Mora-Vallejo, A., Claessens, L., Stoorvogel, J., Heuvelink, G.B.M., 2008. Small scale digital soil mapping in Southeastern Kenya. Catena 76, 44–53. https://doi.org/10.1016/j.catena.2008.09.008

Morellos, A., Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., Mouazen, A.M., 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst. Eng. 152, 104–116. https://doi.org/10.1016/j.biosystemseng.2016.04.018

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

https://doi.org/10.1016/j.geoderma.2013.08.024

Nelson, A., Reuter, H.I., Gessler, P., 2009. Dem Production methods and sources. Dev. Soil Sci. 33, 65–85. https://doi.org/10.1016/S0166-2481(08)00003-2

Odeh, I.O.A., Mcbratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197–214.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100. https://doi.org/10.1016/j.geoderma.2013.09.024

Park, S.J., McSweeney, K.K., Lowery, B.B., 2001. Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma 103, 249–272. https://doi.org/10.1016/S0016-7061(01)00042-8

Paterson, G., Turner, D., Wiese, L., Van Zijl, G., Clarke, C., Van Tol, J., 2015. Spatial soil information in South Africa: Situational analysis, limitations and challenges. S. Afr. J. Sci. 111, 1–7. https://doi.org/10.17159/sajs.2015/20140178

Pienaar, L., Traub, L.N., 2015. Understanding the smallholder farmer in South Africa: Towards a sustainable livelihoods classification, in: International Conference of Agricultural Economists. Milan, p. 36.

Ray, S.S., Singh, J.P., Das, G., Panigrahy, S., 2004. Use of high resolution remote sensing data for generating site-specific soil mangement plan, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Riggers, C., Poeplau, C., Don, A., Bamminger, C., Höper, H., 2019. Multi-model ensemble improved the prediction of trends in soil organic carbon stocks in German croplands. Geoderma 345, 17–30. https://doi.org/10.1016/j.geoderma.2019.03.014

Román Dobarco, M., Arrouays, D., Lagacherie, P., Ciampalini, R., Saby, N.P.A., 2017. Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. Geoderma 298, 67–77. https://doi.org/10.1016/j.geoderma.2017.03.015

Royle, J.A., Nychka, D., 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Comput. Geosci. 24, 479–488.

Scull, P., Franklin, J., Chadwick, O. a., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27, 171–197. https://doi.org/10.1191/0309133303pp366ra

Stum, A.K., 2010. Random forests applied as a soil spatial predictive model in arid Utah. Digit. Soil Mapp. - Bridg. Res. Environ. Appl. Oper. 189. https://doi.org/10.1007/978-90-481-8863-5_15

Triantafilis, J., Laslett, G.M., McBratney, A.B., 2000. Calibrating an electromagnetic induction instrument to measure salinity in soil under irrigated cotton. Soil Sci. Soc. Am. J. 64, 1009–1017.

Triantafilis, J., Lesch, S.M., La Lau, K., Buchanan, S.M., 2009. Field level digital soil mapping of cation exchange capacity using electromagnetic induction and a hierarchical spatial regression model. Aust. J. Soil Res. 47, 651–663. https://doi.org/10.1071/SR08240

van den Bergh, H.M., Weepener, H.L., 2008. Development of Spatial Modelling Methodologies for Semidetailed Soil Mapping, Primarily in Support of Curbing Soil Degradation and the Zoning of High Potential Land. Pretoria, South Africa.

van Zijl, G., 2019. Digital soil mapping approaches to address real world problems in southern Africa. Geoderma 337, 1301–1308. https://doi.org/10.1016/j.geoderma.2018.07.052

van Zijl, G., van Tol, J., Tinnefeld, M., Le Roux, P., 2019. A hillslope based digital soil mapping approach, for hydropedological assessments. Geoderma 354, 113888. https://doi.org/10.1016/j.geoderma.2019.113888

Van Zijl, G.M., Le Roux, P.A., Turner, D.P., 2013. Disaggregation of land types using terrain analysis, expert knowledge and GIS methods. South African J. Plant Soil 30, 123–129. https://doi.org/10.1080/02571862.2013.806679

Vašát, R., Heuvelink, G.B.M., Borůvka, L., 2010. Sampling design optimization for multivariate soil mapping. Geoderma 155, 147–153. https://doi.org/10.1016/j.geoderma.2009.07.005

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2016. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142. https://doi.org/10.1016/j.geoderma.2016.06.006

Wahren, T.F., Julich, S., Nunes, J.P., Gonzalez-Pelayo, O., Hawtree, D., Feger, K.H., Keizer, J.J., 2016. Combining digital soil mapping and hydrological modeling in a data scarce watershed in north-central Portugal. Geoderma 264, 350–362. https://doi.org/10.1016/j.geoderma.2015.08.023

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

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists., 2nd ed, Statistics in Practice. John Wiley & Sons, Inc. https://doi.org/10.2136/vzj2002.0321

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

Wet, P. de, 2018. Only 0.2% of South Africa's farms are bigger than 12,000 hectares – and may qualify for land expropriation [WWW Document]. Bus. Insid. SA. URL https://www.businessinsider.co.za/12000-hectare-farms-are-a-tiny-fraction-in-sa-but-big-agriculture-produce-all-the-food-2018-8

Xu, Y., Smith, S.E., Grunwald, S., Abd-elrahman, A., Wani, S.P., Nair, V.D., 2018. Estimating soil total nitrogen in smallholder farm settings using remote sensing spectral indices and regression kriging. Catena 163, 111–122. https://doi.org/10.1016/j.catena.2017.12.011

Zeraatpisheh, M., Ayoubi, S., Brungard, C., Finke, P., 2019. Disaggregating and updating a legacy soil map using DSMART, fuzzy c-means and k-means clustering algorithms in Central Iran. Geoderma 340, 249–258. https://doi.org/10.1016/j.geoderma.2019.01.005

Zhang, S., Shen, C., Chen, X., Ye, H., Huang, Y., Lai, S., 2013. Spatial Interpolation of Soil Texture Using Compositional Kriging and Regression Kriging with Consideration of the Characteristics of Compositional Data and Environment Variables. J. Integr. Agric. 12, 1673–1683. https://doi.org/10.1016/S2095-3119(13)60395-0

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

Chapter 2 High resolution digital soil mapping of multiple soil properties; an alternative to the traditional field survey?

This chapter is based on the publication Flynn, T., Clercq, W. De, Rozanov, A., Clarke, C., 2019. High-resolution digital soil mapping of multiple soil properties: an alternative to the traditional field survey? *South African Journal of Plant and Soil, 1-11* found in Appendix A.

Abstract:

Spatial information on soil particle size distribution and soil organic carbon (SOC) are important for land use management, environmental models, and policy making. Digital soil mapping techniques can quantitatively predict these soil properties using minimal resources. However, DSM has not been adequately evaluated at the farm-scale. The aim of this study was to optimise the DSM framework to produce farm-scale soil maps for 366 ha in the Sandspruit catchment, Western Cape, South Africa. Four feature selection techniques and eight predictive models were evaluated on their ability to predict particle size distribution and SOC. A boosted linear feature selection produced the highest accuracy for all but one soil property. The top performing predictive models were robust linear models for gravel (ridge regression, RMSE 9.01%, R^2 0.75), sand (support vector machine, RMSE 4.69%, R^2 0.67), clay (quantile regression, RMSE 2.38%, R^2 0.52), and SOC (ridge regression, RMSE 0.19%, R^2 0.41). Random forest was the best predictive model for silt content (RMSE 4.12%, R^2 0.53). This approach appears to be robust for farm-scale soil mapping where the number of observations is often small but high-resolution soil data is required.

2.1 Introduction

Spatial information on soil particle size distribution (PSD) and soil organic carbon (SOC) is increasingly important for land use and environmental management as these properties influence soil functions such as plant available water (Hollis et al., 1977) evaporation (Wetzel and Chang, 1987), soil aggregate stability (Amézketa, 1999), and compaction (Bodman and Constantin, 1965; Soane, 1990). Digital soil mapping offers a way to predict PSD and SOC at a resolution that conventional soil maps cannot achieve.

There are many predictive models such as various types of linear regression (Chagas et al. 2016; McKenzie and Austin 1993), decision trees (Jafari et al., 2014; Moran and Bui 2002; Subburayalu et al., 2014), random forest (Chagas et al., 2016; Hengl et al., 2015; Pahlavan et al., 2014), generalised

additive models (Bishop and Mcbratney 2001), artificial neural networks (Aitkenhead and Coull 2016; Behrens and Förster 2005; Brungard et al., 2015), and fuzzy logic models (De Gruijter et al., 1997; Lagacherie et al., 1997; Qi et al., 2006; Zhu 1997; van Zijl et al., 2013). There has been much research into predictive models, however, many of the above-mentioned studies conclude that there is no single algorithm that predicts all soil properties best. Therefore, many predictive models must be tried to obtain the best accuracy for the soil property of interest (Kuhn and Johnson 2013).

Although there are many predictive models and a common DSM framework, the *scorpan* framework does not explicitly place feature selection into the framework as the method focuses on capturing all the *scorpan* factors. Due to this, feature selection is not a commonly researched aspect in DSM (Behrens et al., 2010) and McBratney et al. (2003) states that there is a need for more research in this regard. Feature selection algorithms (FSAs) produce a set of covariates which correlate to the soil property of interest. This is important to increase interpretability of the model, reduce runtime of the predictive model, reduce multicollinearity, and to increase model accuracy (Guyon and Elisseeff, 2003).

There are three types of FSAs including wrapper, filter, and embedded methods. Wrapper methods use an iterative processes to find the optimal subset of covariates by either adding (forward recursive) or removing covariates (backwards recursive) until a certain performance criteria is met (Kuhn and Johnson, 2013). Filter methods evaluate each covariate independently and select the covariates based on a correlation threshold (Liu and Motoda, 1998). In other words, filter methods will select any covariate which has a strong relationship with the soil attribute of interest. Some predictive models such as decision trees, random forest, support vector machines, and regularised generalised linear models perform feature selection internally (embedded method). There are also data transformation algorithms such as principle component analysis (PCA) which transforms variables into a new orthogonally linear set of data which is uncorrelated.

Many authors note that different and/or additional covariates could have improved the model and, at the farm-scale, not all *scorpan* factors are needed due to low variance in their values (e.g., climate) over short distances. Feature selection algorithms that have been used in DSM include PCA (Hoffmann et al., 2014), ANOVA analysis (Sun et al., 2011), step-wise linear reduction (Mora-Vallejo et al., 2008), univariate and collinear analysis (Kempen et al., 2009) and recursive feature selection

(Brungard et al., 2015). It is hypothesised that optimising and treating feature selection techniques and predictive models simultaneously, is a robust approach suitable for farm-scale soil mapping.

The objective (Framework 1, Objective 1) of this paper was to simultaneously optimise FSAs and predictive models to produce soil maps at the farm scale. This DSM framework can be seen as an adaptation of the *scorpan* method and specifies the evaluation of FSAs within the framework.

2.2 Methods and materials

2.2.1 Site location and soil samples

The research site lies approximately 33°14'51.72"S and 18°8'52.52"E in the middle of the Sandspruit river catchment, Western Cape, South Africa. The catchment location and sample design are shown in Figure 2.1a and Figure 2.1b, respectively. The site was chosen to capture as much landscape variability as possible at the farm-scale. The area covers 366 ha under dry land agriculture with wheat and canola crop rotation. The altitude ranges between 94 m and 220 m above sea level. The predominant geology of the landscape is greywacke, phyllite, and schist of the Moorreesburg, Klipplaat, and Berg River formations of the Malmesbury group. There are silcrete and ferricrete outcrops near the site but not directly on the area of interest. These outcroppings separate the old African surface and the younger dissected land surface at lower altitudes.

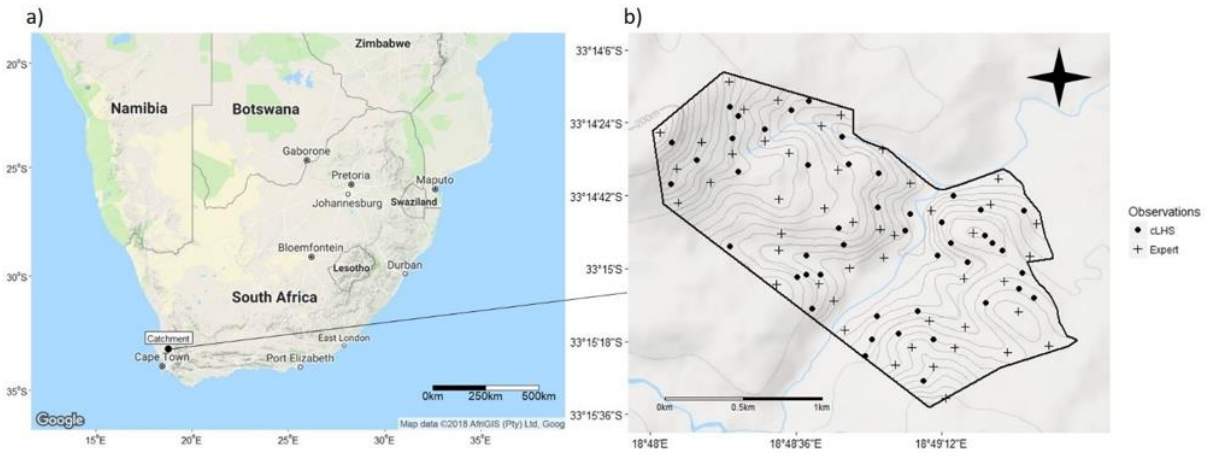


Figure 2.1: The catchment location within South Africa (a) and the site showing the soil sample design (b).

In total, 93 soil profiles were classified and sampled which were used for both training and validating the models. There were two sampling schemes developed to sufficiently cover terrain. The two sampling schemes include 47 systemic soil samples and 46 soil samples using conditioned Latin

hypercube sampling (cLHS). The expert (Dr Freddie Ellis, with 50 years mapping experience in the region) used 5 m contour lines overlaid on a google image to place profiles that would capture maximum soil variation.

Conditioned Latin Hypercube sampling is a type of random sampling strategy that stratifies the samples on the values of multiple covariates (Minasny and McBratney 2006). Conditioned Latin Hypercube uses the probability distribution to form a Latin Hypercube where each row and each column have one sample. In other words, cLHS is an iterative processes to which optimises the stratification from a multivariate distribution. In simplified terms from Minasny and McBratney (2006), for *k* number of iterations, the cLHS processes in this study follows:

1. Take the product of the quantiles of each covariate and *n* strata to get equally. probable intervals. This is to develop a correlation matrix for the covariates (*C*).
2. Then take *n* number of random samples from the covariates to develop a correlation matrix for *x* sampled sites (*T*).
3. Calculate the first objective function (Equation 2.1). Where, η is a matrix of samples *x*, then $\eta\left(q_j^i \leq x_j < q_j^{i+1}\right)$ is the number of samples (x_j) that are in the quantiles q_j^i and q_j^{i+1}.

$$O_1 = \sum_{i}^{n}\sum_{j=1}^{k}\left|\eta\left(q_j^i \leq x_j < q_j^{i+1}\right) - 1\right| \tag{2.1}$$

   - The second objective function is to make sure that the samples follow the distribution of the covariates. The equation for objective function 2 is shown in Equation 2.2. Where, c_{ij} are the elements in *C*, and t_{ij} are the elements in *T*. In other words, the correlation matrix from the original data and the sampled data.

$$O_2 = \sum_{i}^{n}\sum_{j=1}^{k}\left|c_{ij} - t_{ij}\right| \tag{2.2}$$

   - The overall objective function (*O*) is then the sum O_1 and O_2.
4. Calculate (*Metro*) where *Metro* = $[\Delta O/T]^2$ (from simulated annealing).
   - ΔO is the change in the objective function for each iteration.

- T is a number between 0 and 1 known as the cooling temperature.

5. Generate a random number between 0 and 1. If the random number is larger than *Metro* than the sample is removed and replaced.
6. Remove samples with greatest $\eta\left(q_j^i \leq x_j < q_j^{i+1}\right)$ and replace these samples.
7. Repeat the process until either *k* iterations has been reached or *O* meets a stopping criterion.

In this study, cLHS was implemented with 100,000 iterations on six environmental covariates. A large number of iterations were used to make sure the cLHS algorithm sufficiently stratified the feature space. The covariates selected were altitude, slope, aspect, plan curvature, profile curvature, SAGA wetness index (SWI) and a soil adjusted vegetative index (SAVI). These covariates were selected to represent relief and vegetation according to the *scorpan* factors (McBratney et al., 2003). Due to the small area, only relief and vegetation were used as covariates in the cLHS sampling strategy. Additionally, a parent material map was not available at a sufficient detail.

Aspect and slope (calculated in degrees) represent the first derivative of the DEM (see Sec 2.2.2) while plan and profile curvature represent the second derivates perpendicular and parallel to the slope, respectively. These parameters were calculated using the 9 parameter 2nd order polynomial method developed by Zevenbergen and Thorne (1987). The SWI is a compound topographic index that represents places of water accumulation (Conrad et al., 2015). The SWI was calculated using a suction of 10, square root of the catchment area, catchment slope (as opposed to local slope) and no weights. The equation for the SWI can be seen Equation 2.3, where α is the catchment area and β is the catchment slope.

$$SWI = log\left(\frac{\alpha^2}{\tan(\beta)}\right) \tag{2.3}$$

A Sentinel-2A image (band 3 and band 8) was obtained from 23 September 2016 to develop the SAVI. The image was pre-processed into bottom of atmosphere reflectance using Sen2Cor add-on in the Sentinel Application Platform (SNAP; European Space Agency, 2018). Bands 3 and 8 represent red and NIR wavelengths, respectively. A SAVI is similar to a NDVI representing plant vigour, however, it accounts for soil reflectance (Huete, 1988). The equation for the SAVI is shown in Equation 2.4,

where L represents the vegetative cover. The factor L was set to 0.1 as this is suitable for most agriculture fields (Rondeaux et al., 1996).

$$SAVI = \left(\frac{NIR - Red}{NIR + Red + L}\right)(1 + L) \quad (2.4)$$

2.2.2 Primary data sources

A DEM at a 30 m resolution was acquired from the Advanced Land Observation Satellite (ALOS-2) provided by the Japanese Aerospace Exploration Agency (JAXA) http://www.eorc.jaxa.jp/ALOS/en/aw3d30. The DEM was obtained from the S034E018 thumbnail corresponding to the geographic coordinates of 34 degrees south, 18 degrees east. The DEM was selected because the vertical accuracy (6 m) is significantly higher than other freely available DEMs (e.g. 30 m SRTM) at the same resolution. Sentinel-2A images with four spectral bands were used for extraction of indices described below. Four Sentinel-2A images were obtained based on the growth period of wheat in the Western Cape from the 4th February 2017 (fallow), 21st April 2017 (ploughed), 28th August 2016 (growth), and the 23th September 2016 (harvest).

2.2.3 Soil properties

Five topsoil properties were sampled down to the depth of the first subsoil horizon (Albic, B horizon, or lithic contact) to assess the DSM framework. The descriptive statistics and frequency distribution of the soil properties are shown in Table 2.1 and Figure 2.2, respectively. All soil properties were measured after air drying and sieving the soil (<2.0 mm). Gravel content was measured by taking the gravimetric weight percent of the coarse fragments. Sand, silt, and clay content were measured by the pipette method (Gee and Or 2002) and the sand grade was determined by the sieve method (Soil Survey Staff 2014). SOC was measured using the Walkley Black method (Walkley and Black 1934) to determine the soil organic carbon percentage.

Table 2.1: Soil property descriptive statistics showing the mean, standard deviation (Sd), and range.

Soil properties	Mean	Sd	Range
Gravel (%)	39.9	18.3	0.0 - 64.0
Sand (%)	60.3	9.2	19.6 - 76.7
Silt (%)	29.7	7.0	16.5 - 65.6
Clay (%)	10.1	3.7	3.88 - 22.7
SOC (%)	0.65	0.7	0.07 - 1.20

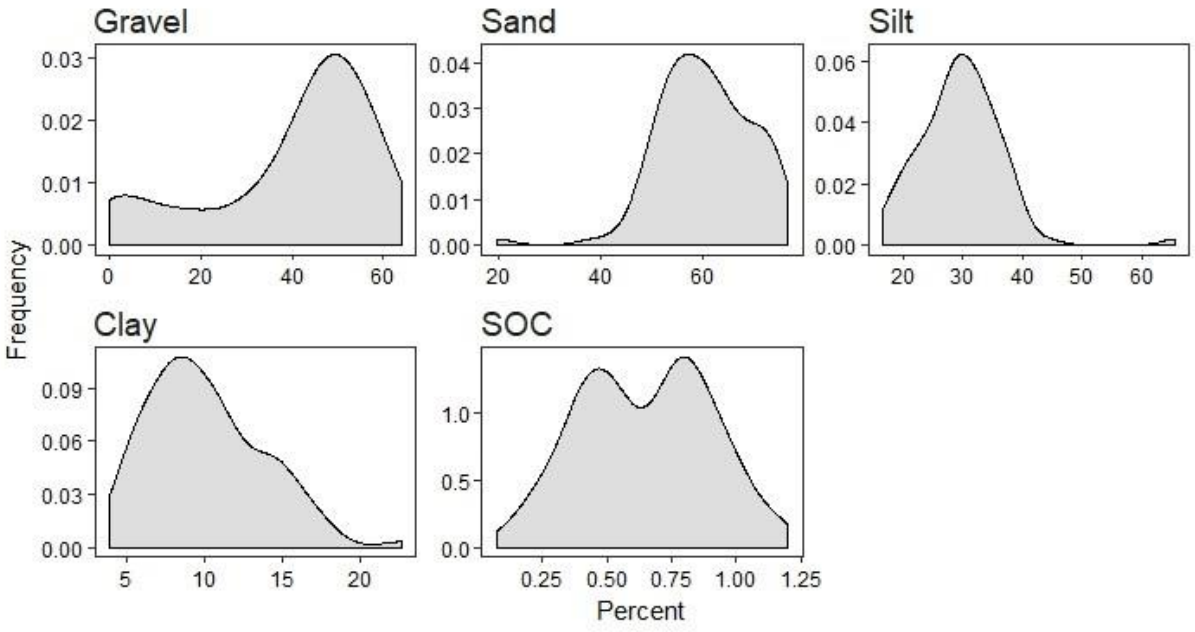


Figure 2.2: Soil property distribution based on 93 soil observations (gaussian kernel density estimation).

2.2.4 Digital soil mapping framework

A diagram of the adapted DSM framework is shown in Figure 2.3. A pool of covariates was developed from which, FSAs selected covariates (see Sec 2.2.6). Each combination of covariates, selected by each FSA, were used to spatially predict each soil property. This was an iterative process for all FSAs and predictive model combinations. In total, 350 models were run as there are 70 possible model combinations for each soil property. The model with the lowest RMSE was considered to be the best performing model. Due to the limited number of samples, the models were validated through leave-one out cross validation (LOOCV). Leave-one out cross validation leaves one observation out, trains the model, and predicts on the single left-out sample. This is done for all observations and the accuracy is averaged.

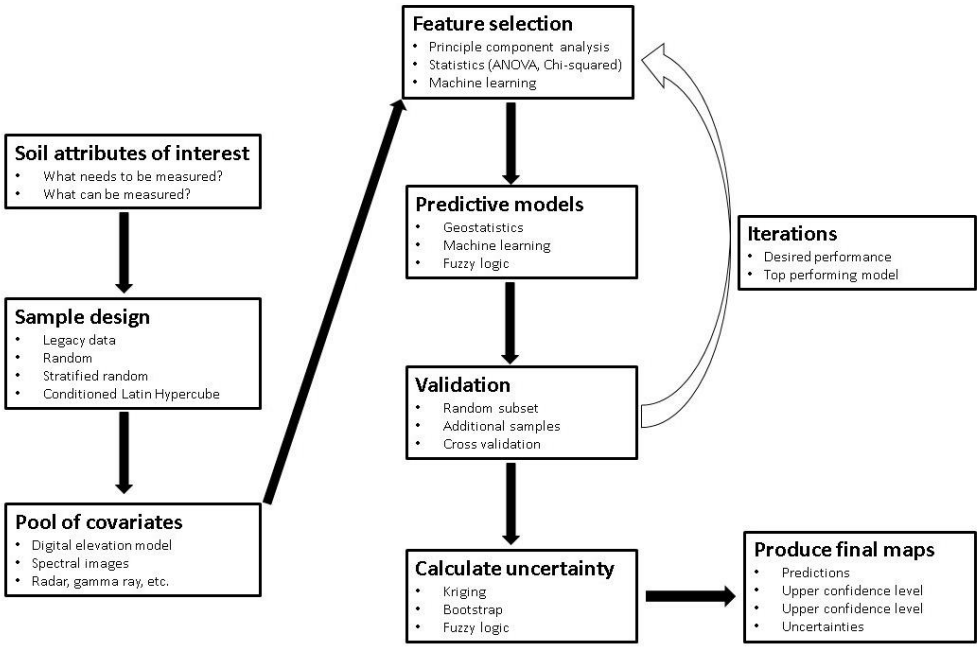


Figure 2.3: Flow diagram of the digital soil mapping methodology.

The mathematical representation of this DSM framework is shown in Equation 2.5. Where, S is the soil property of interest, the function f is a deterministic predictive model which establishes soil environmental relationships with the covariates Q, selected from the feature selection function g, and ε is the independent random error added to the model. The equation assumes that the covariates Q, are developed from all covariates that can be obtained regardless of the expert opinion. In other words, the expert does not know the true relationship between soil properties and the covariates. The $\boldsymbol{\varepsilon}$ term is the residuals from the fitted model which was modelled by a sample variogram. The sample variogram is used to krige the residuals and the values are added to the trend calculated by function f.

$$S = f\big(g(Q)\big) + \varepsilon \tag{2.5}$$

The difference between this method and the *scorpan* method, is quantitative feature selection is specified and optimised together with the predictive model. This approach increases the importance of which covariates are placed into the predictive model as the function f, cannot be defined until the function g, selects appropriate covariates. This specification can help decrease the subjectivity and collinearity of covariate selection when using a large number of covariates (Mason and Perreault

1991). Additionally, this can add to the interpretability of the model as the number of covariates is reduced.

2.2.5 Feature selection

In total, four FSAs were evaluated as shown in Table 2.2. Feature selection algorithms were selected to incorporate a variety of algorithms that have little or no tuning parameters. This includes two linear models and two random forest models. Each FSA was optimised through the LOOCV resampling.

Table 2.2: Feature selection algorithms used in each predictive model.

Technique	Type	Algorithm
Univariate	Filter	Random forest
Recursive	Wrapper	Random forest
LASSO	Embedded	L_1 regularised linear
Boost	Embedded	L_2 regularised linear

Univariate feature selection (UFS) and backwards recursive feature selection (RFS) were implemented in the internal functions of the caret R package through a Random Forest model (Kuhn et al., 2018). The UFS function performs many iterations by trying a random subset of covariates after finding each covariate which correlates with the soil property. The UFS was used because it is an iterative version of filter methods and can handle non-linear relationships. The RFS progressively eliminates covariates until the RMSE either drops or reaches an optimal value. In other words, the RFS algorithm starts with all covariates and progressively eliminates (backwards recursive) unimportant covariates by the integer $2^{2:4}$ from the proceeding set of covariates.

Both algorithms calculate, and rank covariate importance based on the out-of-bag error (OOB) averaged over each tree grown. Out-of-bag errors are calculated by bagging each tree grown. Bagging in RF randomly resamples (without replacement) the 63.2% of the soil observations for each tree. It then predicts over the soil observations left out to determine. The rank of covariate importance is calculated by the average of the decrease in sum of squared errors when a covariate is removed from the model.

A "least absolute shrinkage and selection operator" (LASSO) was implemented through the glmnet R package (Friedman et al., 2010). LASSO is a generalized linear model which minimizes covariate coefficients on the absolute error of the residuals (L_1 regularization). Therefore, if covariates are correlated, the least important covariates coefficient will be minimised to zero deeming the covariate unimportant. The λ value (degree of penalty) was optimised and the coefficients were extracted for the optimal lambda value. The covariates which did not have an absolute value of zero were used in the predictive models. A LASSO feature selection was implemented because LASSO is efficient with high dimensional data sets, improves model interpretation, and does not substantially increase bias (Tibshirani, 1996).

In the case of LASSO, the mathematical representation of L_1 regularisation is shown in Equation 2.6. Where, SSE_{L_1} is the absolute errors, *n* is the number of observations, γ_i is the measured soil property, x_i is the predicted soil value, P is the number of variable, λ is the degree of penalty or in other words, the larger the λ value, the larger the degree the coefficients are minimised, β_j are the estimated coefficients.

$$SSE_{L_1} = \sum_{i=1}^{n}(\gamma_i - x_i)^2 + \lambda \sum_{j=i}^{P}|\beta_j| \tag{2.6}$$

LASSO minimises this function through coordinate descent. The idea is to minimise the loss function one direction at a time. The equation for coordinate descent is shown in Equation 2.7. For each iteration of *k*, x^{k+1} is the next position, x^k is the current position, α_k is the step length (how far the next step will move towards the minima), ∇ is the gradient to local minima, $f(x^k)$ is the SSE_{L_1} at current position of a single coordinate i_k.

$$x^{k+1} = x^k - \alpha_k[\nabla f(x^k)]i_k \tag{2.7}$$

A boosted generalized linear model (Boost) was implemented with the "glmboost" function through the mboost R package (Bühlmann and Hothorn, 2007). The Boost model fits component-wise linear models as base learners and is boosted by correcting for the sum of the squared error (SSE_{L_2}) of the residuals (L_2 regularisation). However, the method of feature selection is a "black box" with little

known on how it selects the covariates. The number of boosts was optimised with pruning. Boost was chosen because it is a novel feature selection technique suitable for high dimensional data (Bühlmann and Hothorn, 2007). Boost feature selection was not implemented for classification models because it only supports binomial classification.

The mathematical representation of regression with L_2 regularisation is shown in Equation 2.8. This equation is essentially the same as equation 2.4, however, instead taking the absolute value of the coefficients, L_2 regularisation takes the sum of the squared coefficients. Therefore, the coefficients cannot be shrunken to zero as opposed to LASSO.

$$SSE_{L_2} = \sum_{i=1}^{n} (\gamma_i - x_i)^2 + \lambda \sum_{j=i}^{P} \beta_j^2 \tag{2.8}$$

Boost minimises the the SSE_{L_1} through gradient descent as opposed to coordinate descent, gradient descent minimises SSE_{L_1} by taking all directions into account simultaneously (i.e., steepest gradient). The equation for gradient descent is shown in Equation 2.9. Where, a_{n-1} is the next position, a_n is the current position, γ is the rate at which the algorithm moves to the minimum, ∇ is the direction of the fastest increase, and $f(a_n)$ is the SSE_{L_2} as current location.

$$a_{n-1} = a_n - \gamma \nabla f(a_n) \tag{2.9}$$

2.2.6 Predictive models

The predictive models implemented are shown in Table 2.3. A general description of each model can be found in Kuhn and Johnson (2013). The models were chosen to get a wide range of machine learning techniques from robust linear to nonlinear tree/rule and additive models. Robust linear models include ridge regression (RR), linear boosted models (LBM), quantile regression (QR) and support vector machines (SVM) with a linear kernel. None-linear models include SVM with a radial kernel, random forest (RF), stochastic gradient boosting (SGB), cubist (CB), and penalised additive spines (P-splines). Model parameters were optimised during the LOOCV resampling.

Table 2.3: Algorithms used for all predictive models.

Algorithm	Model type
General additive with splines (P-splines)	Generalized additive
Stochastic gradient boosting (SGB)	Tree based additive
Ridge regression (RR)	Regularised linear
Linear boosted regression (LBM)	Linear additive
Quantile regression (QR)	Linear
Support vector machine (SVM)	Linear and radial
Cubist	Rule based
Random Forest (RF)	Tree based

Three linear models were implemented in this study. A ridge regression (RR) was implemented through the glmnet R package (Friedman et al., 2010). A ridge regression implements a generalized linear model with L_2 regularization through coordinate descent. Linear boosted models (LBM) were implemented through the mboost R package (Bühlmann and Hothorn, 2007). The LBM model also implements L_2 regularization where the number of boosts were optimised with pruning. Boosting is an ensemble method performed in sequence (additive) to correct for errors while pruning reduces the number of boosts to prevent overfitting. Quantile regression with L_1 regularization on the median was implemented in the quantreg R package (Koenker, 2019). Regression on the median is a robust linear model which does not assume a normal distribution of a soil property.

Random forest (RF) is an ensemble algorithm which grows decision trees in parallel (Breiman, 2001). It does so by bagging the soil observations and growing a user defined number of trees (*ntree)* and user defined number of randomly chosen covariates at each split (*mtry*). Essentially, it is double random making the trees uncorrelated which results in a high bias but low variance model (Hastie et al., 2009). The final predictions are the mean of the ensemble of decision trees.

Random forest models were conducted in the randomForest R package (Breiman, 2001). The *mtry* parameter was optimised and the number of *ntree* was held constant at 1000 trees. The number of trees were held at 1000 because Breiman (2002), states that at least 1000 trees are required for a stable variable importance measure. Random forest was used because it is suitable for small and large data, can handle non-linear relationships, and is robust against over fitting (Breiman, 2001).

Cubist models were implemented in the cubist R package (Quinlan, 1993). Cubist is similar to a decision tree and has the option to be boosted. However, a linear model is performed in each

terminal node and each split. This allows CB models to generalise better than a decision tree. The tree and linear models are then simplified into a set of rules where each rule is associated with a linear model. There were two tuning parameters that were optimised in the CB models. The number of committees is the number of trees grown in sequence (like boosting). The number of neighbours is the number of *k*-nearest neighbours that were used to correct for errors. Cubist was selected because CB is a complex yet an interpretable model as the output of the model defines each rule made.

Stochastic gradient boosting (SGB) was implemented in the gbm R package (Friedman, 2002, 2001). Stochastic gradient boosting is a type of ensemble method which creates decision trees in sequence as opposed to random forest. Stochastic gradient boosting builds a decision tree which produce residuals and the next decision tree randomly samples the residuals of the proceeding tree to correct for errors. The SGB algorithm reduces the sum of squared errors through gradient decent (Friedman, 2001). In contrast to gradient descent, SGB randomly samples each boost by a user ratio of samples (bag fraction) for each step. This also creates a doubly random model like RF and decorrelates each tree.

The SGB algorithm has five main parameters. The learning rate was held constant at 0.01, minimum number of observations in each terminal node was set to 10, and the bag fraction was set to 0.5. However, the number of trees grown, and number of interactions was optimised. A SGB was used because it represents an alternative to random forest and has been shown to have similar performance by (Forkuor et al., 2017; Hitziger and Ließ, 2014).

A boosted generalised additive model with L_2 penalized splines (P-spline) was implemented through the mboost R package (Bühlmann and Hothorn, 2007). Splines are defined by a stage wise polynomial function that acts as a smoothing base learner for each covariate in the model. The model is then boosted (additive) on the residuals through gradient descent to create the final model. The number of boosts was optimised with pruning, knots were set to 20, and degrees of freedom set to four. Knots are the number of places the spline from each covariate meet. The knots and degrees of freedom values were set based on the recommendations by Bühlmann and Hothorn (2007). P-spline was implemented because it is a novel boosting algorithm which has performed well in the Kaggle competition (Taieb and Hyndman, 2013).

A linear and radial kernel support vector machine (SVM) was implemented through the Kernlab R package. Support vector machines project the soil samples into higher dimensions using a kernel function (Drucker et al., 1996). Kernel functions are a type of function that allows the soil property values to be projected into higher dimensions to separate the soil properties more efficiently. The easiest way to display this is through a linear kernel shown in Equation 2.10. Where, *K* is the kernel function, x, y are vectors of n-dimensions, *f* is a function that projects the values into m-dimensions where, m-dimensions > n-dimensions.

$$K(x, y) = < f(x), f(y) > \tag{2.10}$$

In the SVM models, the cost function (coefficient penalty) and σ values were both optimised. The cost is a threshold or regularisation technique that eliminates outliers from the regression model. The σ value determines the width of the gaussian distribution for SVM with a radial kernel. For example, with high σ values, the boundaries from the support vectors is determined only by the values closest to these boundaries. When σ is low, soil samples further away from the closest values are considered. Support vector machines were used because they are known for being one of the best "out of the box" classifiers and have been adapted as a robust regression algorithm. They are also suitable with high dimensional data set where there are more covariates than soil observations.

2.2.7 Covariate development

A pool of covariates was developed to capture organisms, relief, parent material, and age according to the *scorpan* factors. The covariates were easily calculated topographic derivatives, and spectral images that were thought to capture the environmental variation sufficiently. In total, 47 topographic covariates and 36 spectral covariates were developed. Many covariates (relative to soil observations), were used to decrease subjectivity in covariate selection and evaluate feature selection techniques.

The covariates were resampled to 10 m using a bi-cubic spline. This resolution corresponds to the finest legible resolution according to inspection density for 366 ha and 93 soil observations (Hengl, 2006). The concept of inspection density is that maps are predicted from point data and therefore, should have a similar sample density per area. The equation for the finest legible resolution can be seen in Equation 2.11. Where A is the area of the site in square meters (3,660,000 m^2) and N is the

number of soil observations (93 observations). The equation for the finest legible resolution is shown in Equation 2.2, where A is the square meters of the area and N is the number of soil observations.

$$\text{Finest resolution} = 0.05 \times \sqrt{\frac{A}{N}} \tag{2.11}$$

Topographic covariates are shown in Table 2.4. A description of all topographic covariates can be found in Hengl and Reuter (2009). Covariates that might be redundant (e.g. standardised height and normalised height) were not removed as to evaluate feature selection techniques. All spectral covariates, a description of the covariates, and their associated calculations are listed in Table 2.5. A description of the spectral covariates can be found in Huete (1988), and Ray et al., (2004). The spectral bands and indices were selected to incorporate soil, age, parent material, and vegetation according to the *scorpan* method.

Table 2.4: All topographic covariates derived from the ALOS-2 DEM and their SAGA module representation.

Representation	Covariate
Land form elements	Land type geomorphon (LTS-GM)
Hydrology characteristics	Channel network base level
	Flow directions
	Catchment area
	Catchment slope
	Flow path length
	Modified catchment area
	Slope length
	Slope length factor (LS factor)
	Stream power index
	SAGA wetness index (SWI)
	Topographic wetness index
	Melton ruggedness index
Lighting/exposure	Analytical Hillshading
	Diffuse insolation
	Direct insolation
	Negative openness
	Positive openness
	Sky view factor
	Visible sky
Local morphometry	Flow line curvature
	Plan curvature
	Profile curvature
	Tangential curvature
	Total curvature
	Aspect (degrees)
	Convergence Index
	Convexity
	Cross section curvature
	Elevation
	General curvature
	Downslope gradient
	Longitudinal curvature
	Mass balance index
	Maximum curvature
	Slope (degrees)
	Terrain ruggedness index
	Vector ruggedness index (VRI)
Landscape morphometry	Maximum height
	Mid-slope position
	Multiresolution ridge top flatness index (MRRTF)
	Multiresolution valley bottom flatness index (MRVBF)
	Normalized height
	Slope height
	Standardized height
	Topographic position index
	Valley depth

Table 2.5: Soil and vegetative bands and indices used for soil spatial variability analysis.

Bands	Band Length (μm)	Symbol	Resolution
Red	0.665	Red	10 m
Near infrared (NIR)	0.842	NIR	10 m
Short wave infrared 1	1.610	SSWIR	10 – 20 m
Short wave infrared 2	2.190	LSWIR	10 – 20 m

Indices	Calculation	Property	Resolution
Brightness Index	$\frac{R^2 + G^2 + B^2}{3^{0.5}}$	Average reflectance	10 m
Coloration Index	$(R - G)/(R + G)$	Soil colour	10 m
Redness Index (RI)	$R\text{^}2/(B * G\text{^}3)$	Hematite	10 m
Saturation Index (SI)	$(R - B)/(R + B)$	Spectral slope	10 m
Soil adjusted vegetative index	$\frac{NIR - R}{NIR + R + 0.1}(1 + 0.1)$	Chlorophyll reflectance	10 m

2.2.8 Spatial autocorrelation

Spatial autocorrelation was evaluated through residual variogram analysis. Due to the small sample size, only isotropic variograms were analysed as there were not enough point-pairs for an anisotropic variogram (Webster and Oliver 2001). The variograms were estimated through residual maximum likelihood analysis (REML) as this is considered best practice (Lark et al., 2006). REML performs general least squares regression but corrects the residuals by maximising the log-likelihood measure. The benefit of REML is the variance term does not need to be stationary and the covariates can be correlated in space and time. Additionally, a variogram fitted through REML needs less samples than other least squares methods (Kerry and Oliver 2007). Residuals of each model were evaluated for normality before variogram analysis and the kriged residuals were added back to the trend of the deterministic model.

2.2.9 Spatial uncertainties

Prediction ranges were estimated by developing fuzzy *k*-means clusters with extragrades (FKMe) of covariates with similar distribution of model errors (McBratney and de Gruijter 1992). The prediction range is a measure of uncertainty that gives the range of uncertainty of predictions at a given pixel for a specific confidence interval. For example, if the confidence interval is 0.9, the prediction range of sand is 5%, and the mean is 30% than there is a 90% chance that the predictions fall between 25% and 35% for a cluster. In this paper, 0.9 confidence intervals were used.

The FKMe algorithm is similar to fuzzy *k*-means, however, it creates a fuzzy cluster which incorporates observations that are far from fuzzy member centroids (averages). These clusters are often areas with a high uncertainty of model predictions. The approach used here follows Malone et al., (2011), who used fuzzy membership to classify environmental covariates with a similar distribution of model errors. In each fuzzy cluster, a prediction range is created by taking the weighted mean of the model errors. In other words, the covariates are clustered based on the confidence interval (0.9) of the model errors. A prediction interval was then developed for each cluster based on the distribution of the model errors. The prediction range is then calculated based on its membership to each cluster. This final product is the prediction range in each cluster.

2.3 Results and discussion

2.3.1 Optimised models

In this study, many predictive models were run because it cannot be assumed that a model will outperform another (Kuhn and Johnson 2013; Wolpert et al., 1996). The strategy here was to try a number of models and focus on the top performing model. Each model was optimised through the LOOCV procedure using the default settings in the caret R package (Kuhn et al., 2018).

The results for the top performing feature selection and predictive model combinations are shown in Table 2.6. The predictive model (Ridge regression) for both gravel and SOC had an optimal λ value of 0.01, which is the degree by which the model panelises covariates. The sand SVM model had a linear kernel with a cost function of one. The silt RF model *mtry* parameter was optimised with 16 covariates. According to the out of bag error (OOB), the RF model explained 48% of the variance of silt which corresponds well with the validation results before regression kriging. This indicates that the internal evaluation in the RF model is a good indication of model performance. Clay predictions were best estimated by a QR on the median with no tuning parameters.

Table 2.6: Results for the top performing model for each soil property.

Property	Model	Krige	Selection	RMSE	R^2
Gravel	RR	Yes	Boost	9.01%	0.75
Sand	SVM	Yes	Boost	4.69%	0.67
Silt	RF	Yes	RFS	4.12%	0.53
Clay	QR	Yes	Boost	2.38%	0.52
SOC	RR	Yes	Boost	0.19%	0.41

Satisfactory results were achieved for gravel and sand content in terms of RMSE and R^2 values. Both property predictions are comparable to other studies. For example, Ballabio et al., (2016), achieved an RMSE of 19.22% and an R^2 of 0.73 for gravel content. The authors used multivariate adaptive regression splines to map coarse fragments on the continental scale. In Burkina Faso, Forkuor et al., (2017) achieved an internal validation RMSE of 7.59% and R^2 of 0.34 for sand content using SVM. However, the authors achieved a more satisfactory result using multilinear regression. Silt, clay, and SOC have less satisfactory R^2 values, however, the RMSE values are similar to other studies. In Kenya, Mutuma et al., (2016) used RF to predicted silt and clay content with a RMSE of 7.30% and 9.90%, respectively. In Mozambique, Cambule et al., (2013), mapped SOC with a RMSE 0.42% through kriging with external drift.

These results show a trend towards Boost feature selection with robust linear models. This suggests that this combination can be used as a powerful alternative to more complex models when predicting soil properties from a small data set. An explanation for gravel and SOC results is that the L_2 regularisation is suitable for small data sets (Kuhn and Johnson 2013) and slightly increases the bias to lower the variance (Hastie et al., 2009). For sand, the SVM with a linear kernel is suitable for high dimensional data sets (Drucker et al., 1996). While a quantile regression accounts for the skewed clay distribution and therefore, quantile regression does not need normalised data (Koenker et al., 1978). Alternatively, these results may be a result of complex models over fitting the small data set (Hastie et al., 2009).

2.3.2 Final predictions

Final predictions for all soil properties are shown in Figure 2.4. The final maps were discretised into three prediction quantiles. The legend for each soil property can be seen in Table 2.7. This was to simplify the maps to increase interpretability. Gravel, silt, and SOC spatial predictions appear realistic and mirror what was visually observed in the field. The maps produced from the optimisation procedure for sand and clay were significantly affected by spectral images on different fields, causing discrete and unrealistic boundaries. These boundaries were seen after regression kriging. When spectral images were removed, the maps appeared more realistic. However, the RMSE for sand increased by 0.92% and the R^2 decreased by 14%. For clay, the RMSE increased by 0.31% and the R^2 decreased by 11%. Due to the more realistic maps produced without spectral covariates, both sand

and clay predictions had spectral covariates removed. These results suggest that the top performing model (FSA + predictions) may not be the optimal model to produce a farm-scale soil map. Therefore, each map should be inspected visually on pedological knowledge in addition to statistical evaluation. It is recommended that the simplest most realistic model be used as the final product.

Preferential erosion has removed the finer particles on crest and mid-slope positions resulting in the residual accumulation of gravel upslope (Shi and Schulin, 2018). On the other hand, fluvial sands have resulted in the absolute accumulation of sand on lower elevations and therefore, sand has an increasing trend downslope. Soils at higher elevations developed from residual highly weathered material of the old African surface (Lambrechts, 1983; Scholms1983). Soils developed from this parent material, have a higher iron content than the younger soils below which could be preventing the removal of finer particles by stabilising the soil aggregates (Barral et al., 1998). The higher clay content on upslope positions may also be protecting the SOC through sorption and micro-aggregation (Singh et al., 2017). Therefore, upslope positions have a higher silt, clay, and SOC content.

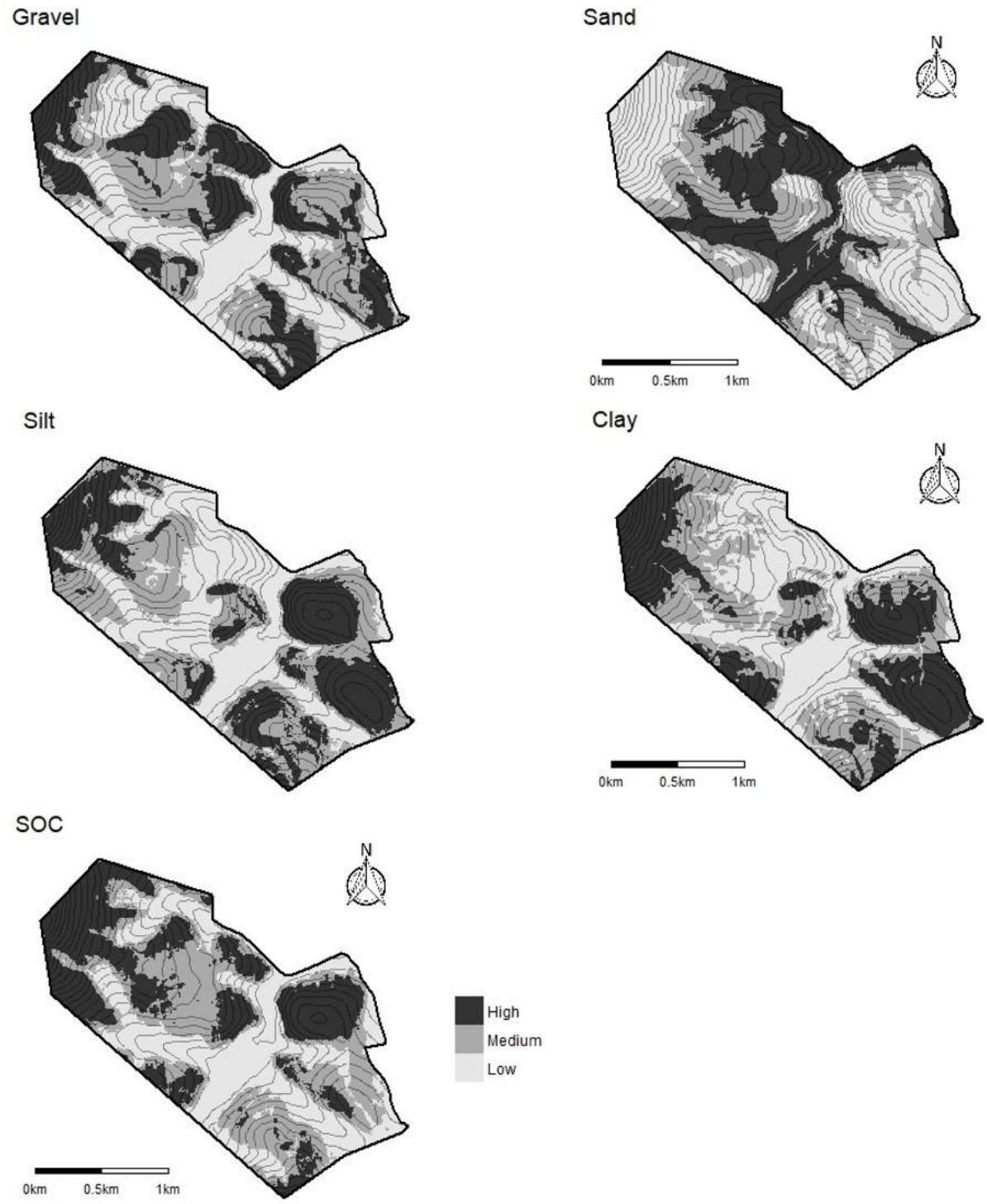


Figure 2.4: Prediction for gravel, sand, silt, clay, and SOC content.

Table 2.7: Legend for soil property prediction rasters.

Property	Low		Medium		High	
	Min%	Max%	Min%	Max%	Min%	Max%
Gravel	0.00	39	39	49	49	70
Sand	37.1	56.5	56.5	63.4	63.4	78.4
Silt	19.5	27.1	27.1	31.0	31.0	41.3
Clay	3.88	8.93	8.93	11.5	11.5	22.6
SOC	0.01	0.58	0.58	0.71	0.71	1.03

2.3.3 Covariate importance

The five most important covariates for each soil property are shown in Table 2.8. The covariate importance for gravel, sand, clay, and SOC is the scaled absolute value of the linear coefficients. For example, in a simple linear model where $Y = -3X + 2$, the absolute coefficient is 3. The values are then scaled from 0 to 100%, where the further away from zero the coefficient, the higher the importance. The covariate importance for silt is based on the mean squared error or in other words, the larger the increase in the mean squared error when the covariate is removed, the larger the importance of that covariate.

Table 2.8: Top five most important covariates for each soil property where rank, represents the order of importance.

Rank	Gravel	Sand	Silt	Clay	SOC
1	Mid-slope Position	Flow Path Length	Flow Path Length	RI fallow	Convexity
2	LSWIR ploughed	SWI	SWI	Total Curvature	Normalized height
3	LS Factor	Negative Openness	Normalized Height	Flow Line Curvature	SSWIR ploughed
4	Normalized Height	Mid-slope Position	Negative Openness	Negative Openness	Slope height
5	VRI	VRI	Convexity	SI growing	Aspect

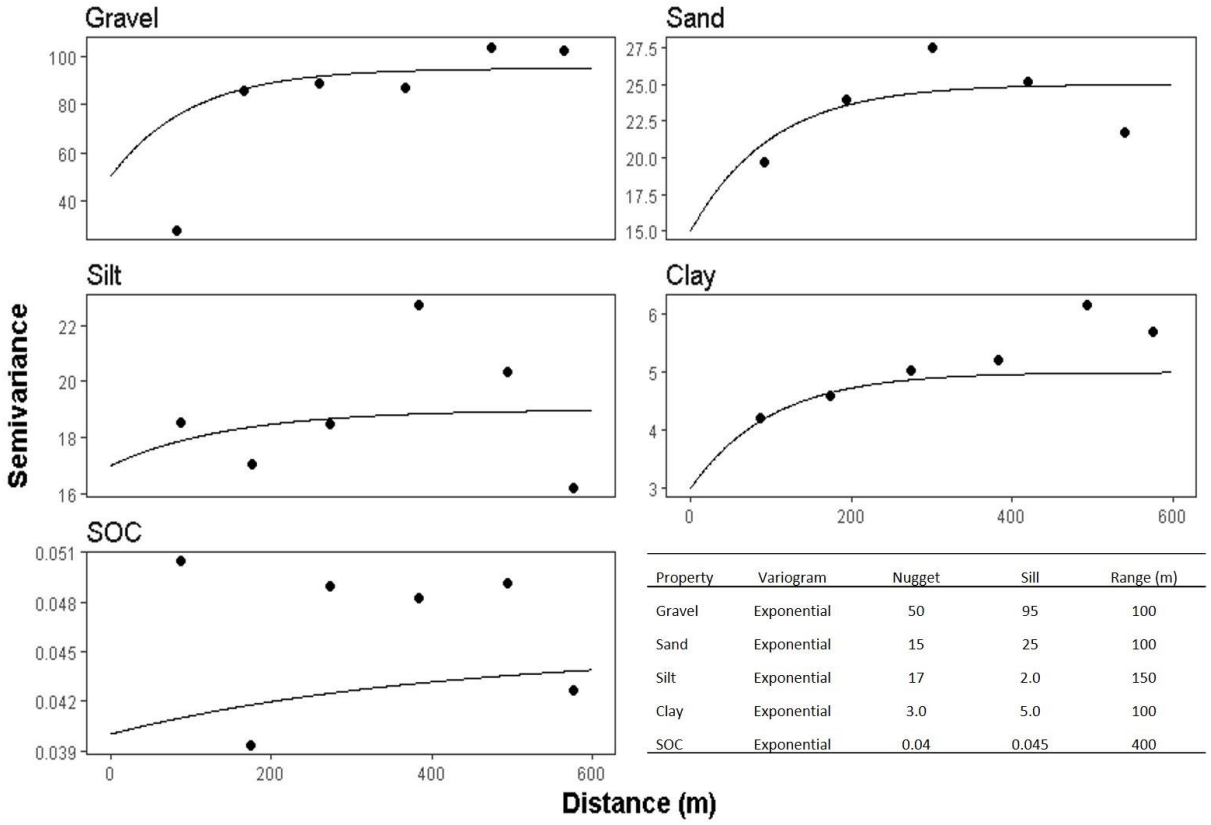
The number of covariates used to predict gravel, sand, silt, clay, and SOC was 17, 12, 16, 19, and 14, respectively. For each soil property, topographic covariates were used in a larger proportion than spectral covariates. Sand and clay were the only property, which heavily relied on spectral covariates. Overall the most important covariates were associated with slope position and shape, soil moisture, and solar radiation. Therefore, topography is influencing soil distribution through erosion and depositional processes as well as solar radiation (Beaudette and O'Geen 2009; Brungard et al., 2015).

Additionally, a multitemporal approach is important when using spectral covariates. For example, a RI (fallow) and a SI (growth) are among the top five important covariates for clay. These relationships would have been difficult to detect from expert knowledge alone and this justifies the feature selection approach.

2.3.4 Spatial autocorrelation

The sample variogram and variogram parameters for each soil property are shown in Figure 2.5. It should be noted that the soil properties did not need to be transformed as all residuals had a normal distribution according to a shapiro normality test ($p<0.05$). Spatial prediction of all the soil properties improved with regression kriging, however this improvement was relatively small. However, this indicates that the models did not capture all of the soil variability in the predictive models which show bias and can be misleading. The RMSE of gravel, sand, silt, clay, and SOC improved by 0.11%, 0.15%, 0.17%, 0.14%, and 0.007%, respectively. Perhaps the largest improvement was seen in the R^2 values of clay and SOC, which improved by 7% and 5%, respectively. The sample variogram for gravel, sand, and clay showed the most spatial autocorrelation and have the most reliable predictions.

The variogram for silt and SOC have a low nugget to sill ratio and may be unreliable. This can be attributed to the variograms being estimated from 93 soil observations which is below the minimum sample size recommended by Oliver and Webster (2014). Additionally, the lack of spatial autocorrelation could be due to the sample design. For example, over 75% of the soil samples are over 300 m apart with an average spacing of 763 m. Therefore, the processes determining the spatial distribution of SOC are acting on a scale that might not be represented on this site and/or from the soil sample design.



Property	Variogram	Nugget	Sill	Range (m)
Gravel	Exponential	50	95	100
Sand	Exponential	15	25	100
Silt	Exponential	17	2.0	150
Clay	Exponential	3.0	5.0	100
SOC	Exponential	0.04	0.045	400

Figure 2.5: Residual variogram for all top performing regression models.

2.3.5 Spatial uncertainties

To evaluate soil property uncertainties, prediction ranges were created through FKMe. Figure 2.6 and Table 2.9 show the prediction range and descriptive statistics for each soil property, respectively. The extragrades cluster for most of the soil properties, correspond to what was reported by McBratney and de Gruijter (1992) as the extragrades cluster encompass places of high uncertainty. However, the extragrades cluster for gravel has the lowest uncertainty. Furthermore, the extragrades cluster represents places where there is no soil such as the stream or places with dense bush which were not sampled. Therefore, care is recommended when interpreting this cluster. Besides the extragrades cluster, places of highest uncertainty are associated with concave slopes.

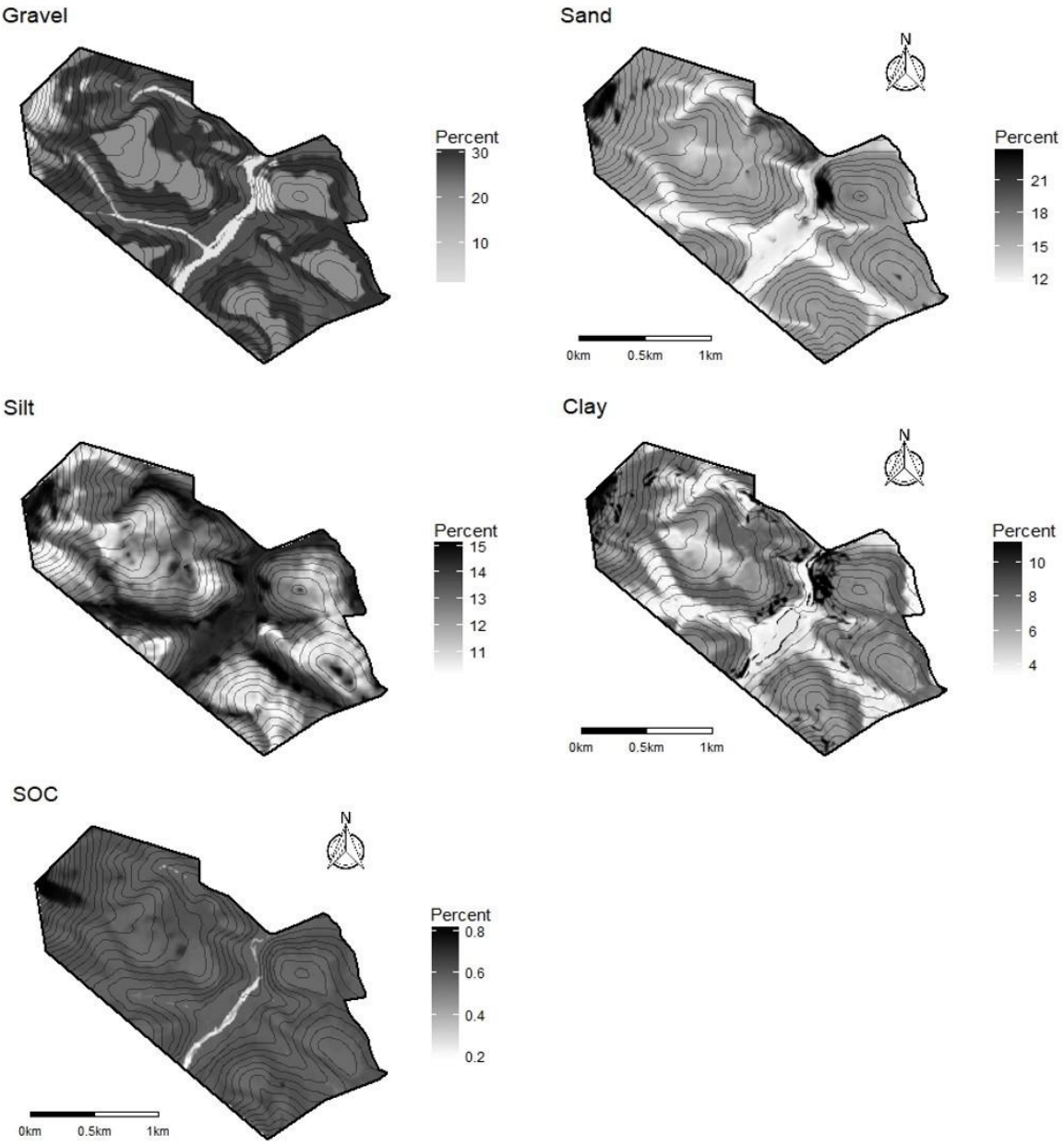


Figure 2.6: The range (%) for gravel, sand, silt, clay, and SOC predictions.

Table 2.9: Descriptive statistics for prediction range of each soil property.

Property	Clusters	Min (%)	Mean (%)	Max (%)
Gravel	4	0.0	25	30
Sand	3	11.6	15.8	23.7
Silt	3	10.2	12.8	15.1
Clay	3	3.40	6.74	11.3
SOC	3	0.00	0.63	1.06

2.4 Conclusion

This study has shown that feature selection predictive model optimisation is effective at predicting gravel, sand, silt, clay, and SOC at the farm-scale in the Sandspruit catchment. The models were evaluated on their RMSE and R^2 for each individual soil property. The spatial uncertainties were evaluated through prediction ranges developed through FKMe. This approach is important where financial resources are low but high-resolution soil data is required because it requires less field work than a traditional soil survey. The conclusions of this study are:

- The simultaneous optimisation of feature selection and predictive models proves to be a robust approach to predict soil properties at the farm-scale. However, the maps still need to be evaluated on pedological knowledge.
- Boost feature selection and robust linear models obtained the highest end accuracy for four out of five soil properties.
- Regression kriging increased the accuracy for all soil property predictions.
- Spatial uncertainties suggest the highest uncertainty is associated with slopes (as opposed to hill tops and valleys).

This research lays out a DSM methodology and in theory, the methodology can be applied across South Africa at a detailed scale. This can create cost efficient soil maps in different regions of South Africa at a scale suitable for land use management on smallholder farms. However, a cost benefit analysis is required to determine the actual financial cost reduction of this methodology. Additionally, this framework needs to be evaluated in different geographic areas and at different scales.

2.5 References

Aitkenhead, M.J., Coull, M.C., 2016. Mapping soil carbon stocks across Scotland using a neural network model. Geoderma 262, 187–198. https://doi.org/10.1016/j.geoderma.2015.08.034

Amézketa, E., 1999. Soil aggregate stability: A review. J. Sustain. Agric. 14, 83–151. https://doi.org/https://doi.org/10.1300/J064v14n02_08

Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123. https://doi.org/10.1016/j.geoderma.2015.07.006

Barral, M.T., Arias, M., Guerif, J., 1998. Effects of iron and organic matter on the porosity and structural stability of soil aggregates. Soil Tillage Res. 46, 261–272.

Beaudette, D.E., O'Geen, A.T., 2009. Quantifying the Aspect Effect: An Application of Solar Radiation Modeling for Soil Survey. Soil Sci. Soc. Am. J. 73, 1345. https://doi.org/10.2136/sssaj2008.0229

Behrens, T., Förster, H., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 25, 580–591. https://doi.org/10.1002/jpin.200421414

Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155, 175–185. https://doi.org/10.1016/j.geoderma.2009.07.010

Bishop, T.F.A., Mcbratney, A.B., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103.

Bodman, G.B., Constantin, G.K., 1965. Influence of Particle Size Distribution . in Soil Compaction. Hilgardia 36, 567–592.

Breiman, L., 2002. Manual on setting up, using, and understanding random forests v3.1, Technical Report, Statistics Department University of California Berkeley. https://doi.org/10.2776/85168

Breiman, L., 2001. Random Forests. Berkeley, California. https://doi.org/10.1017/CBO9781107415324.004

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

Bühlmann, P., Hothorn, T., 2007. Boosting Algorithms: Regularization, Prediction and Model Fitting. Stat. Sci. 22, 477–505. https://doi.org/10.1214/07-STS242

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013. A methodology for digital soil mapping in poorly-accessible areas. Geoderma 192, 341–353. https://doi.org/10.1016/j.geoderma.2012.08.020

Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena 139, 232–240. https://doi.org/10.1016/j.catena.2016.01.001

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analysis (SAGA). Geoscientific Model Development. https://doi.org/doi:10.5194/gmd-8-1991-2015

De Gruijter, J.J., Walvoort, D.J.J., Van Gaans, P.F.M., 1997. Continuous soil maps - A fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma 77, 169–195. https://doi.org/10.1016/S0016-7061(97)00021-9

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1996. Linear Support Vector Regression Machines, in: Advances in Neural Information Processing Systems. MIT press.

European Space Agency, 2018. SNAP. Sentin. Appl. Platf. v6.0.0.

Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M., 2017. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. PLoS One 12. https://doi.org/10.1371/journal.pone.0170478

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1–24. https://doi.org/10.18637/jss.v033.i01

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anlysis 38, 367–378.

Friedman, J.H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. 29, 1189–1232.

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. Springer Series in Statistics.

Hengl, T., 2006. Finding the right pixel size. Comput. Geosci. 32, 1283–1298. https://doi.org/10.1016/j.cageo.2005.11.008

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., De Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One 10, 1–26. https://doi.org/10.1371/journal.pone.0125814

Hengl, T., Reuter, H.I., 2009. Geomorphology: Concepts, Software, Applications. Dev. Soil Sci. 33.

Hitziger, M., Ließ, M., 2014. Comparison of Three Supervised Learning Methods for Digital Soil Mapping: Application to a Complex Terrain in the Ecuadorian Andes. Appl. Environ. Soil Sci. 12.

Hoffmann, U., Hoffmann, T., Jurasinski, G., Glatzel, S., Kuhn, N.J., 2014. Assessing the spatial variability of soil organic carbon stocks in an alpine setting (Grindelwald, Swiss Alps). Geoderma 232–234, 270–283. https://doi.org/10.1016/j.geoderma.2014.04.038

Hollis, J.M., Jones, R.J.A., Palmer, R.C., 1977. The effects of organic matter and particle size on the water-retention properties of some soils in the west midlands of England. Geoderma 17, 225–238. https://doi.org/10.1016/0016-7061(77)90053-2

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25, 295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Jafari, A., Khademi, H., Finke, P.A., Van de Wauw, J., Ayoubi, S., 2014. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. Geoderma 232–234, 148–163. https://doi.org/10.1016/j.geoderma.2014.04.029

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach, Geoderma. Elsevier B.V. https://doi.org/10.1016/j.geoderma.2009.04.023

Kerry, R., Oliver, M.A., 2007. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. Geoderma 140, 383–396. https://doi.org/10.1016/j.geoderma.2007.04.019

Koenker, R., 2019. Quantreg: Quantile Regression. R Packag. version 5.35.

Koenker, R., Bassett, G., Jan, N., 1978. Regression Quantiles. Econometrica 46, 33–50.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, the R.C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., 2018. Caret: Classification and Regression Training.

Lagacherie, P., Cazemier, D.R., Van Gaans, P.F.M., Burrough, P.A., 1997. Fuzzy k-means clustering of fields in an elementary catchment and extrapolation to a larger area. Geoderma 77, 197–216. https://doi.org/10.1016/S0016-7061(97)00022-0

Lambrechts, J.J.N., 1983. Soils, Soil Process and Distribution in the Fynbos Region: An Introduction. Council for Scientific and Industral Research, Pretoria.

Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 97, 787–799. https://doi.org/10.1111/j.1365-2389.2005.00768.x

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Springer Science+Business Media, LLC, New York.

Malone, B.P., Mcbratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160, 614–626.

https://doi.org/10.1016/j.geoderma.2010.11.013

Mason, C.H., Perreault, W.D., 1991. Collinearity, Power, and Interpretation of Multiple Regression Analysis. J. Mark. Res. 28, 268–280.

McBratney, A.B., de Gruijter, J.J., 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades. J. Soil Sci. 43, 159–175. https://doi.org/10.1111/j.1365-2389.1992.tb00127.x

McBratney, A.B., Santos, M.L.M., Minasnya, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

McKenzie, N.J., Austin, M.P., 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57, 329–355. https://doi.org/10.1016/0016-7061(93)90049-Q

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

Mora-Vallejo, A., Claessens, L., Stoorvogel, J., Heuvelink, G.B.M., 2008. Small scale digital soil mapping in Southeastern Kenya. Catena 76, 44–53. https://doi.org/10.1016/j.catena.2008.09.008

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. Int. J. Geogr. Inf. Sci. 16, 533–549. https://doi.org/10.1080/13658810210138715

Mutuma, E., Csorba, A., Michéli, E., 2016. Prediction of soil properties using Mid-Infrared Spectroscopy and Random Forest regression in the Eastern slopes of Mt . Kenya Region. Agric. Sci. Res. J. 6, 253–262.

Oliver, M.A., Webster, R., 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. Catena 113, 56–69. https://doi.org/10.1016/j.catena.2013.09.006

Pahlavan Rad, M.R., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 232–234, 97–106. https://doi.org/10.1016/j.geoderma.2014.04.036

Qi, F., Zhu, A.X., Harrower, M., Burt, J.E., 2006. Fuzzy soil mapping based on prototype category theory. Geoderma 136, 774–787. https://doi.org/10.1016/j.geoderma.2006.06.001

Quinlan, J.R., 1993. Combining Instance-Based and Model-Based Learning. Mach. Learn. 76, 236–243.

Ray, S.S., Singh, J.P., Das, G., Panigrahy, S., Group, A.R., Centre, S.A., Potato, C., 2004. Use of High Resolution Remote Sensing Data For Generatin Site-Specific Soil Mangement Plan. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 127–131.

Rondeaux, G., Steven, M., Baret, F., 1996. Optimization of soil-adjusted vegetation indices. Remote Sens. Environ. 55, 95–107. https://doi.org/10.1016/0034-4257(95)00186-7

Scholms, B.H.A., Ellis, F., Lambrechts, J.J.N., 1983. Soils of the Cape Coastal Platform, in: Deacon, H.J., Hendey, Q.B., Lambrechts, J.J.N. (Eds.), . Council for Scientific and Industral Research, Pretoria.

Shi, P., Schulin, R., 2018. Erosion-induced losses of carbon, nitrogen, phosphorus and heavy metals from agricultural soils of contrasting organic matter management. Sci. Total Environ. 618, 210–218. https://doi.org/10.1016/j.scitotenv.2017.11.060

Singh, M., Sarkar, B., Sarkar, S., Churchman, J., Bolan, N., Mandal, S., Menon, M., Purakayastha, T.J., Beerling, D.J., 2017. Stabilization of Soil Organic Carbon as Influenced by Clay Mineralogy Stabilization of Soil Organic Carbon as Influenced by Clay Mineralogy. Adv. Agron. 148, 1–51. https://doi.org/10.1016/bs.agron.2017.11.001

Soane, B.D., 1990. The role of organic matter in soil compactibility: A review of some practical aspects. Soil Tillage Res. 16, 179–201. https://doi.org/10.1016/0167-1987(90)90029-D

Soil Survey Staff, 2014. Keys to soil taxonomy, 12th ed. USDA-Natural Resources Conservation Service, Washington, DC.

Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334–345. https://doi.org/10.1016/j.geoderma.2013.08.018

Sun, X.L., Zhao, Y.G., Zhang, G.L., Wu, S.C., Man, Y.B., Wong, M.H., 2011. Application of a Digital Soil Mapping Method in Producing Soil Orders on Mountain Areas of Hong Kong Based on Legacy Soil Data. Pedosphere 21, 339–350. https://doi.org/10.1016/S1002-0160(11)60134-3

Taieb, S. Ben, Hyndman, R.J., 2013. A gradient boosting approach to the Kaggle load forecasting competition. Int. J. Forecast. 1–19.

Tibshirani, R., 1996. Regression Selection and Shrinkage via the Lasso. J. R. Stat. Soc. 58, 267–288. https://doi.org/10.2307/2346178

Walkley, A., Black, I.A., 1934. An examination of Degtjareff method for determining soil organic matter, and proposed modification of the chromic acid tritation method. Soil Sci. 37, 29–38.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists., 2nd ed, Statistics in Practice. John Wiley & Sons, Inc. https://doi.org/10.2136/vzj2002.0321

Webster, R., Oliver, M.A., 2001. Geostatistics for Environmental Scientists, in: Statistics in Practice. Wiley, Chichester.

Wetzel, P.J., Chang, J.-T., 1987. Concerning the relationship between evapotranspiration and soil moisture. J. Clim. Appl. Meteorol. https://doi.org/10.1175/1520-0450(1987)026<0018:CTRBEA>2.0.CO;2

Wolpert, D.H., Macready, W.G., 1996. No Free Lunch Theorems for Optimization. IEEE Trans. Evol. Comput. 1, 1–32.

Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative anlsysis of land surface topography. Earth Surf. Process. Landforms 12, 47–56.

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

Chapter 3 Farm scale soil patterns derived from automated terrain classification

This chapter is based on a paper (accepted for publication) Flynn, T., de Clercq, W., Rozanov, A., Clarke, C., Farm-scale soil patterns derived from automated terrain classification. *Catena*.

Abstract:

Landform elements (LFEs) are commonly used in soil science to demarcate pedological boundaries and as a first indication of soil spatial variability. A novel LFE classification system known as geomorphons, has been shown to be able to overcome limitations of other automated LFE classifiers. The pattern recognition algorithm classifies the 10 most common LFEs, is computationally efficient, and is robust to changes in scale. However, due to their novelty, research into geomorphons has been limited. This study aimed to stratify the soil landscape through an aggregated geomorphon at the farm-scale (1:25 000) in the Western Cape, South Africa (33.25° S and 18.20° E). Twenty-four geomorphons were created at different resolution and their association with soil classes were compared. The best fitting geomorphon was aggregated into a 5-unit system corresponding to the South African national resource inventory. The aggregation was based on a decision tree corresponding to soil type. The 5-unit system was evaluated on how well the system stratified soil associations, soil lightness, soil electrical conductivity (EC), soil organic carbon, effective rooting depth (ERD), depth to lithology, gravel, sand, silt, and clay. The prediction potential was compared between the original geomorphon, the aggregated geomorphon, and a manually delineated LFE system. It was found that the aggregated geomorphon stratified all soil attributes except EC. Additionally, the aggregated geomorphon predicted 6 out of 9 soil properties with the lowest RMSE. This study shows that aggregating geomorphons can stratify the soil landscape even at the farm-scale and can be used as an initial indication of the soil spatial variability. This has implications in resource poor areas where an additional soil survey is not feasible or can be used to aid in the disaggregation of existing soil-terrain datasets.

3.1 Introduction

Landforms are commonly used to delineate polypedons because landform boundaries separate pedological and hydrological processes such as accumulation, deposition, and leaching potential (Evans, 2012a). On the hillslope scale, landforms are divided into LFEs which are places with similar shape, gradient, aspect, moisture regime, and/or landscape position (MacMillan and Shary, 2009).

From these boundaries, the spatial variability of soils can be approximated and insight into pedogenic processes can be obtained from either expert knowledge (Botha, 2016) and/or DSM techniques (Odeh et al., 1994). Additionally, landform elements can help produce soil sampling designs for DSM (van Zijl et al., 2019) and to represent relief according to the *scorpan* framework (McBratney et al., 2003). Therefore, quantifying LFEs has the ability to improve soil map accuracy while maintaining relationships that are easily recognised by soil scientists (Silva et al., 2016).

Traditionally, 2-dimensional LFEs have been manually delineated from topographic sheets. Manually delineated systems were commonly used in soil legacy datasets, such as LTS (Land Type Survey Staff, 1972 - 2006), the SOTER (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012), and the U.S. soil survey (Soil Survey Staff, 2017). Early studies into soil-landform relationship include Bleeker and Speight (1978), determined the association between soil type and manually delineated LFEs in Papua New Guinea. The authors note that only a generalised realisation of the soil distribution can be established through LFEs. In New Zealand Tonkin and Basher (1990), used a conceptual landform model to determine the erodibility of soils in the Southern Alps. However, manually delineated models are time consuming to produce and will differ between soil scientists as their delineation is a subjective process. (Jasiewicz et al., 2014). Therefore, manually delineated LFEs are difficult to reproduce as their delineation is a subjective process.

Perhaps the first 3-dimensional semi-automated LFE classification system was produced by Pennock et al. (1987). The authors developed a set of rules from slope, vertical curvature, and horizontal curvature to produce LFEs based on Ruhe's (1960) classification system from digital topographic data. These LFEs consisted of convex shoulders, concave foot-slopes, linear back-slopes, sloping summits and toe-slopes. The use of horizontal curvature was a novel implementation to account for lateral hydrological movement (convergence vs. divergence), an aspect that traditional 2-deminsional systems do not account for (Huggett, 1975; Pennock et al., 1987).

More recent LFE classification systems utilise unsupervised learning algorithms on DEMs and/or DEM derivatives (e.g., slope, curvature) within, or not in, a given distance from a central point (Jasiewicz et al., 2014). Commonly used systems include clustering algorithms such as Self-Organizing Data Analysis Technique (Zhong et al., 2009), fuzzy k-means clustering (Irvin et al., 1997; Schmidt and Hewitt, 2004), and nested-means (Iwahashi and Pike, 2007). Other systems include the incorporation

of contextual information (surrounding environment) as in object based classification (Drăgut and Eisank, 2012) and the use of multiple moving windows in the topographic position index classification (Jenness, 2006; Weiss, 2000). These LFE classification systems have the advantage of being reproducible and quantitative. However, these systems can be time consuming to compute, can be subject to geographic scale, and/or do not fit the conceptual pedological model (landscape position).

Developed by Jasiewicz and Stepinski (2013), geomorphons are a pattern recognition algorithm that classifies the ten most common LFEs from a single pass of a DEM. These LFEs include flat, peak, ridge, shoulder, spur, slope, hollow, foot-slope, valley, and pit which are classified through an automatically adjusting moving window (search radii). According to the authors, geomorphons are computationally efficient, are flexible in terms of scale, and can account for landscape position. Although geomorphons help diminish traditional limitations, DEM derivatives such as geomorphons still need to be evaluated in South Africa in terms of resolution and accuracy (Atkinson et al., 2017).

Research into geomorphons includes Kramm et al. (2017), who found that geomorphons had a substantially higher accuracy with the ground truth at finer resolutions (5 and 10 m) than at coarser resolutions (30 m). Libohova et al. (2016), predicted geomorphons through morphological soil properties using multinomial regression and linear discriminatory analysis. The authors found that aggregating geomorphons into a 5-unit unit system increased the accuracy of predictions. Silva et al. (2016), determined the most suitable geomorphons for two different watersheds in Brazil using three different resolutions and differing search radii. However, the authors found that geomorphons did not stratify the landscape well in terms of particle size distribution in one of the catchments. Pinto et al. (2016), mapped soil water transmissivity through fuzzy logic, environmental covariates, and geomorphons. The authors found that using geomorphons with fuzzy logic had a superior accuracy when compared with the Iwahashi and Pike approach (Iwahashi and Pike, 2007).

Although there are many automated LFE classification systems, relatively little research has been conducted at the farm-scale (>1:25,000) even though LFE processes are affecting soil formation at this scale and hence, the soil spatial variability. Additionally, aggregating LFEs that coincide with soil legacy data to improve existing data has generally been a subjective process based on expert opinion. The objective of this study was (Framework 1, Objective 2) to quantitatively select the appropriate geomorphon and aggregate the geomorphon units to stratify and predict multiple soil

attributes. This study presents a methodology which can be applied to additional areas or combined with national soil databases to increase the detail on a local scale. This is notable in resource poor areas such as Southern Africa where soil information is scarce or unobtainable through conventional soil surveys.

3.2 Methods and materials

3.2.1 Study site

The research area was shown in Chapter 1. The Sandspruit, which drains in a generally north-easterly direction towards the larger Berg River, was responsible for incision of a former more extensive plain, thought to have formed during Early Miocene. Partridge and Maud (1987), reported this former surface as related to the "African Surface" they identified in many places in South and Southern Africa. Since the Mesozoic, deep weathered profiles and duricrust cappings (laterite, silcrete and calcrete) now shape the landscape after various uplift, planation and downcutting cycles. In the Sandspruit region, only remnants of laterite and silcrete were found in the 130 – 150 m above sea-level positions. The mere existence of these materials complicated their classification and to predict soil patterns in the area, as these soils represent multiple cycles of erosion, deposition, and/or pedogenesis (Guillocheau et al., 2018).

3.2.2 Soil classification and properties

Soils were classified for each soil profile according to the Soil African Soil Classification System (Soil Classification Working Group, 1991) and were reclassified into the United States Department of Agriculture (USDA) Soil Taxonomy (Soil Survey Staff, 2014) down to a depth of 1.5 m. The soils were aggregated into six soil associations based on the dominant soil morphological properties as seen in the field. The aggregation was necessary because some soil types such as Aquic Palexeralfs, were only observed once. The soil associations classified, and their description are shown in Table 3.1. These groupings are common practice when using the South African Soil Classification System.

Table 3.1: Soil associations classified, count, USDA equivalence, and their descriptions.

Association	Count	Symbol	Subgroups	Description
Apedal	13	Ap	Typic Haploxerept	Apedal soil structure through 75 cm of soil profile.
Aquic	11	Aq	Typic Epi/Endoaqualf Typic Endoaquept Plinthic Haplaquox	Hydromorphic features present in the soil profile starting within 50 cm of soil surface.
Duplex	10	Dp	Typic Haploxeralfs Aquic Palexeralfs	Strong structured B horizon within 50 cm of soil surface.
Lithic dry	30	Ld	Lithic Haploxerept Lithic Xerorthents	Lithic contact within 50 cm of soil surface with no signs of wetness within 150 cm of soil surface.
Lithic wet	21	Lw	Aquic Haploxerept Aquic Xerothents	Lithic contact within 50 cm of soil surface with signs of wetness within 150 cm of soil surface.
Hard plinthic	8	Hp	Aeric Plinthoaquox Plinthic Petraquents	Hard plinthic layer (relic or in phase) thicker than 5 cm within 50 cm of soil surface.

For each soil profile, the topsoil was sampled (depth of horizon), airdried and sieved (<2 mm). Soil properties measured were colour (lightness), soil EC, soil organic carbon (SOC), effective rooting depth (ERD), depth to lithology (DL), gravel, sand, silt, and clay percent. The density distribution of the measured soil properties is shown in Figure 3.1. The density distribution is based on the gaussian kernel distribution which is a probability distribution of a random variable. Soil colour measurements were conducted using a Konica Minolta CM-600d spectrophotometer (Minolta, Osaka, Japan). Soil colour was recorded in the L*a*b* colour space, where L represents lightness (where 0 is black and 100 is white), a* represents the red-green axis, and b* represents the blue-yellow axis. To capture the soil bleaching phenomenon, only the L value was used for analysis. Electrical conductivity was measured using a glass electrode (Jenway 4510) in a 1:2.5 soil to water ratio. Particle size distribution and SOC measurements were described in Chapter 2.

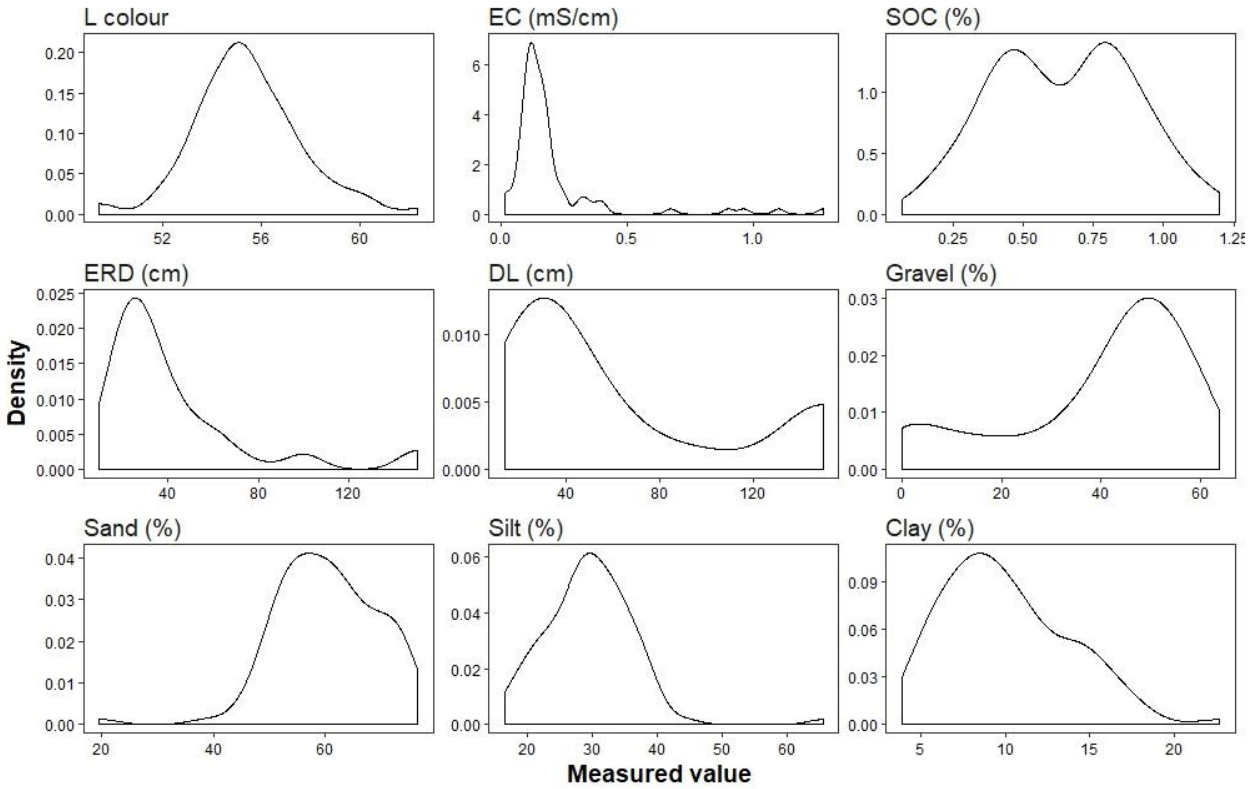


Figure 3.1: Density distribution of measured soil properties (gaussian kernel density estimation).

3.2.3 *Stratification summary*

The soil-landscape stratification processes is shown in Figure 3.2. First, 24 geomorphons were derived. The best fitting geomorphon was selected based on a bootstrapped Cramer's V (CV) test to determine the geomorphons association with soil classes. The selected geomorphon was aggregated into units that resemble TMUs through a decision tree. The estimated TMUs were evaluated on their ability to stratify the soil-landscape using a generalised least squares model that accounted for spatial-autocorrelation. Finally, bootstrap analysis was used to see how well the geomorphon predicts soil properties. The predictions were compared against the GM-10 and a manually delineated LFE classification.

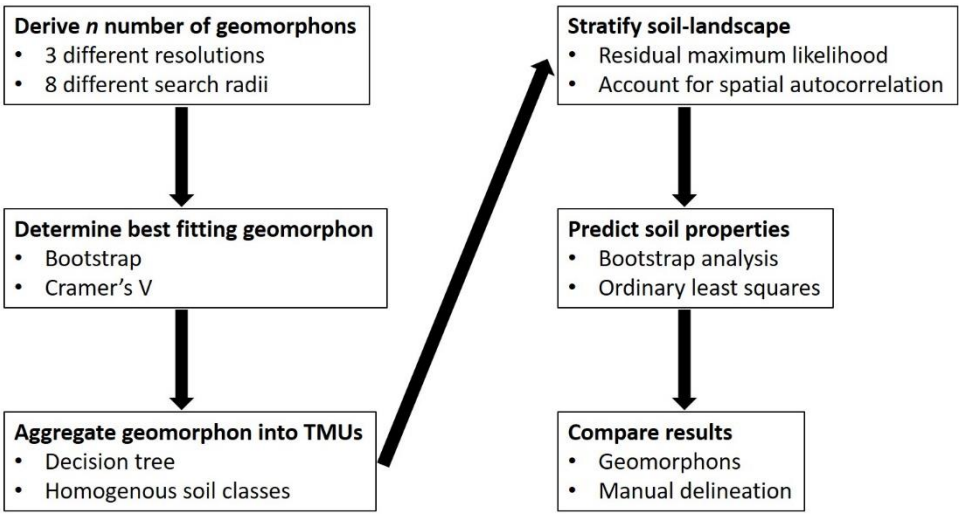


Figure 3.2: Workflow of stratification process from geomorphon selection, aggregation, soil-landscape stratification, soil property predictions, and evaluation.

3.2.4 Landform element grids

Geomorphons utilise a "line-of-sight" approach, relating two different angles over eight directions from a central point. The two angles are known as the zenith and nadir angles. The zenith angle is the angle between 90 degrees (overhead) and the line of sight (0 degrees). The nadir angle is the angle from -90 degrees (below) to the line of sight. The user can specify two parameters in the algorithm, the search radius and flatness threshold. The search radius is the radius the algorithm will search away from a central point to distinguish landscape patterns. The flatness threshold defines what is and is not considered flat. A graphical representation of the LFE classified through the geomorphon algorithm can be seen in Figure 3.3.

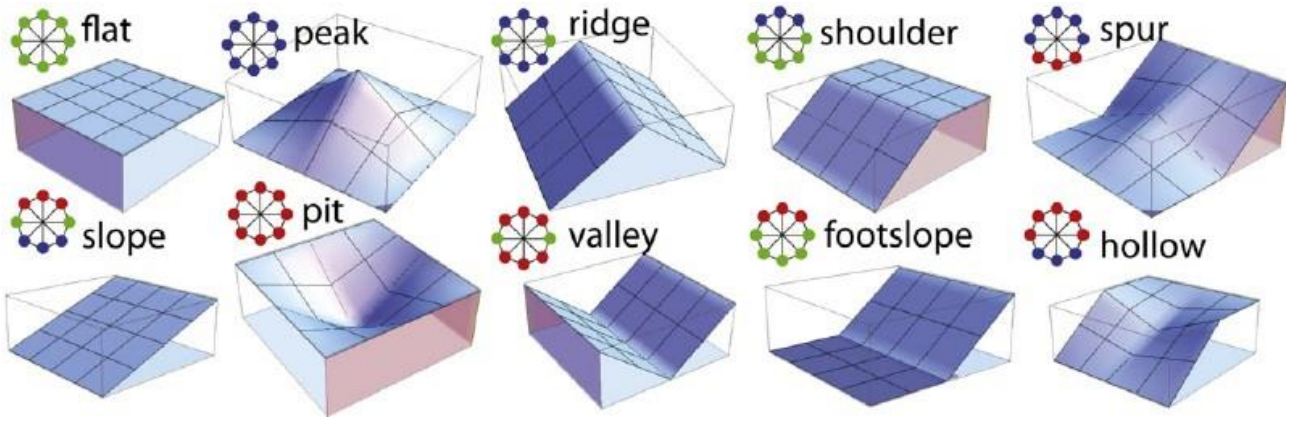


Figure 3.3: The ten most common landform elements classified by geomorphons (Jasiewicz and Stepinski, 2013).

The DEM was loaded into GRASS GIS (Geographic Resources Analysis Support System; GRASS Development Team, 2017) where geomorphons were classified using the r.geomorphons add-on developed by Jasiewicz and Stepinski (2013). Twenty-four geomorphons were derived with a search radius of 20, 25, 30, 40, 50, 60, 100, and 200 cell search radii at a 10, 20, and 30 m resolution. Initially, the geomorphons were developed for the whole Sandspruit catchment but were cropped down to the study area. Therefore, the geomorphons considers the landscape which surrounds the research area.

Smaller search radii were not included due clear discrepancies in the classification (e.g., sloping positions in valleys) and due to a clear pixilation effect. It was thought that these issues made geomorphons with a lower search radii a poor representation of the site. Additionally, the geomorphon algorithm has been shown to produce more stable results at search radii above 30 cells (Di Stefano and Mayer, 2018). The inner search radius was left as default (0 cells) as it was seen that increasing this value substantially decreased computational efficiency.

Out of the 24 geomorphons, the best fitting geomorphon was taken for further analysis. It was decided that the best fitting geomorphon be used instead of a multi-resolution approach because using one realisation of the LFE distribution is more easily interpretable to farmers. The best fitting geomorphon was determined through a bootstrapped Cramer's V test (CV). This is similar to the approach by Silva et al. (2016), who used a χ^2 test of independence to determine the best fitting geomorphon, however, the CV determines strength of association (Liebetrau, 1983). The CV was conducted with 2000 iterations to determine each geomorphons association with the soil classes. A large number of iterations were used to get 99% confidence intervals and get a reliable estimate of the mean CV for all interactions.

The equation for CV can be seen in Equation 3.1 where, *n* is the number of observations and therefore, *n* is 93. The *k* variable is the number of rows or in other words, the number of soil associations. Therefore, *k* equals six because the soils were aggregated into six associations.

$$CV = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{3.1}$$

3.2.5 Terrain morphological units

To correspond with the LTS, the best fitting geomorphon (GM-10) was aggregated into a 5-unit geomorphon (GM-5) corresponding to TMUs. To aggregate the GM-10 into GM-5, a decision tree was used to determine which units showed a difference in the distribution of soil associations. GM-10 units which did not show a statistically different soil distribution, were aggregated into the same group.

To determine if the geomorphon algorithm can mimic the "mental model" used to define TMUs in the LTS, an additional manually delineated TMUs was created (expert LFE). The expert LFE was developed by Dr Freddie Ellis. The expert LFE classification was created by delineating boundaries from a Google Earth satellite image overlaid with 5 m contour lines. The delineation of the expert GM was based on elevation, slope curvatures, and landscape position.

Decision trees all have a common framework where the covariate data is split to increase the homogeneity of the soil property of interest. This is done recursively until the data can no longer be split. The result is a prediction of the distribution of soil associations in the terminal node. Decision trees were applied because they can mimic the "mental model" of a soil scientist and the structure is easily interpreted (Bui and Moran, 2001). For this study, the splits were based on chi-squared test of independence as implemented in the party package (Hothorn et al., 2006).

3.2.6 Soil property stratification

To infer difference in mean soil properties between GM-5 units, a residual maximum likelihood model (REML) was implemented. Residual maximum likelihood models estimate the mean through ordinary least squares; however, the spatial correlation and variance are accounted for by introducing a variogram and variance function on the residuals, respectively. REML is known for being the best unbiased estimate for spatial data and does not require a randomised sample design (Lark et al., 2006).

All models were automatically fit with a variogram in the nlme R package (Pinheiro et al., 2018). REML fits a variogram model and kriges the residuals within each group. The variogram equation is shown in Equation 3.2, where $\gamma(\mathrm{h})$ is the semivariance of the soil property at lag distance h, E is the estimated variance by averaging all point pairs for lag distance h, $Z(x)$ is the measured soil property

at location *x*, and $Z(x+h)$ is the soil property at location (x + h). As can be seen in the variogram equation, the semivariance at lag *h*, depends on distance alone (Cressie, 1985).

$$\gamma(\mathrm{h}) = \frac{1}{2}\mathrm{E}[(\mathrm{Z}(\mathrm{x}) - \mathrm{Z}(\mathrm{x}+\mathrm{h}))^2] \tag{3.2}$$

Each model was selected by comparing the Akaike Information Criteria (AIC) values for different variance functions. The variance function is estimated through maximising the log-likelihood of the residuals within each GM-5 unit. This is an iterative process which stops when the set of variance components (random effects) have the highest likelihood. The variance functions compared includes identity (none), exponential, and power functions. The residuals were checked for normality and homogeneity. If the residuals were not normal, a log transformation was performed. The difference between GM-5 units was compared using a Tukey-Kramer post hoc test ($p < 0.05$).

3.2.7 Soil property predictions

The best fitting GM-10 and its aggregated GM-5 as well as the expert LFE were compared by predicting each soil property through bootstrap resampling. Bootstrapping takes random samples (with replacement) and predicts over the soil samples. This model takes 93 random samples (the same observation can be selected), predicts the soil property over the soil samples, and the RMSE is measured for each resample. The RMSE is taken as the empirical distribution of the prediction accuracy from which, quantiles can be measured. Essentially, it produces more data than the number of soil samples by randomly sampling for a user defined number of resamples. Another benefit of bootstrap resampling is it's a non-parametric method which creates a normal distribution through averaging the accuracy measure (Efron and Tibshirani, 1986).

Bootstrap resamples were used because of the low number of observations and because bootstrap resamples can bypass the assumptions of linear models such as normal distribution of the data (Firdaus et al., 2012). Bootstraps were performed with 2000 iterations for each soil property to get a stable estimate of the average RMSE and to make sure the RMSE values fall within a normal distribution. Since bootstrapping bypasses assumptions of ordinary least squares, a simple linear regression was implemented to predict soil properties. A post hoc Student T test was used to determine the differences in predictions between the LFE systems ($p < 0.05$).

3.3 Results and discussion

3.3.1 Best fitting geomorphon

To determine the best fitting GM-10, the association between soil associations and each geomorphon was evaluated. The association with soil associations was used because soil classification is interpretable to soil scientists (Minasny and McBratney, 2007) and should hold more information than specific soil properties alone (Moore and Russell, 1966). Therefore, it is hypothesised that geomorphons with the highest association, will also correlate well to additional soil attributes such as specific soil properties.

The mean CV values for the bootstrap resamples is shown in Figure 3.4. For all resolutions, as search radius increases so does the association between geomorphons and soil classes. Additionally, resolution strongly affects the CV values at lower search radii. As the search radii increases, the resolution becomes less important. This suggests that at larger radii, geomorphons become more homogenous between 10, 20, and 30 m resolutions. The 10 m resolution geomorphons show the largest change of CV values with an increase in search radius while the 20 m resolution shows the least change. An explanation for this is that a 20 m resolution roughly corresponds to Tobler's (1987) rules (~ 25 m) and is near the optimal resolution for this area.

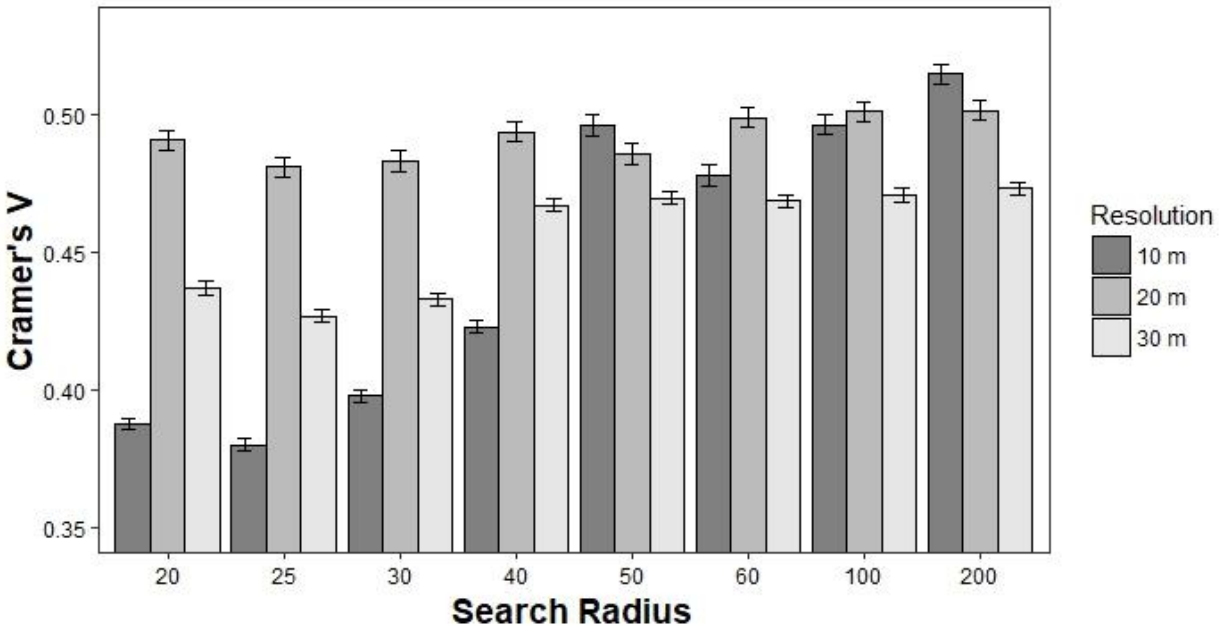


Figure 3.4: Bootstrapped Cramer's V testing the association between geomorphons with 8 different search radius at a 10, 20, and 30 m resolution with six soil classes.

The geomorphon with a 10 m resolution and a 200-cell search radius is the best fitting geomorphon with an average CV of 0.52. This geomorphon statistically outperformed the other geomorphons in their respective groups according to the bootstrap resamples ($p < 0.05$). The best fitting geomorphon was expected for two reasons: i) the smooth topography lacked an abundance of microfeatures (e.g., hollow) and therefore was better represented by a large search radius which smoothed these features into macro features (e.g., slope), ii) the rolling topography in the area indicates that the landscape is defined by more convexity and concavity. As the search radius increases, geomorphons represent convexity and concave landscapes better (Silva et al., 2016). These findings correspond to Kramm et al. (2017), who found geomorphons with a finer resolution correspond more to ground observations and to Zhang and Montgomery (1994), who found a 10 m resolution better represents surface processes than a 30 m resolution. However, both Silva et al. (2016), and Roecker and Thompson (2010), found a higher association with soil classes at a coarser resolutions.

The best fitting geomorphon (GM-10) and the expert LFE estimates are shown in Figure 3.5. The GM-10 did not classify all 10 LFEs in the area. Instead, only peak (PK), ridge (RI), spur (SP), slope (SL), hollow (HL), valley (VL), and pit (PT) were classified. Nevertheless, no geomorphon classified all 10 units on this site due to the undulating topography. This was expected because of the small area and therefore, the absence of some LFEs should be assumed at the farm-scale. As found in previous studies, sloping positions were classified with the highest abundance on the site (Libohova et al., 2016; Silva et al., 2016). However, the area of slope was more predominant for the GM-10 (34%) than the expert LFE (29%). The high presence of sloping positions can be attributed to the Sandspruit cutting into the older planar surface (African erosional surface) creating the present undulated landscape.

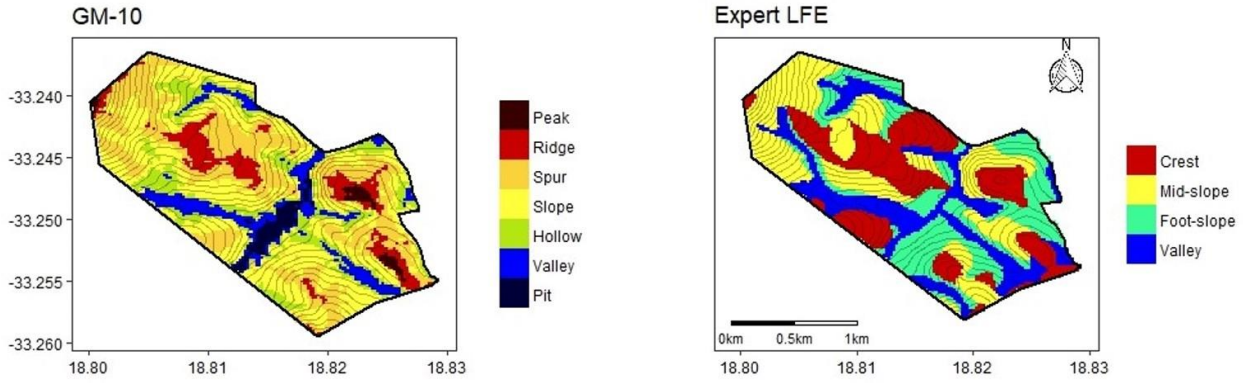


Figure 3.5: Landform element distribution of the GM-10 and expert LFE overlaid on 5 m contours.

3.3.2 Aggregated geomorphon

A decision tree was implemented to determine which GM-10 units should be aggregated to obtain the GM-5. This process also serves to determine if the GM-10 and GM-5 are stratifying the soil-landscape into a more homogenous soil distribution. The decision tree for the GM-10 is shown in Figure 3.6. According to the decision tree, peak, ridge, and spurs are similar, and valleys and pits have a similar soil distribution. Therefore, node 2 was aggregated into valleys, node 5 into mid-slopes, node 6 into crests, and node 7 into foot-slopes. As expected, there is no soil association which is incorporated into one LFE unit. However, the soil distribution between each unit was significantly different ($p < 0.05$). It should be noted the decision tree splits are not shown for the GM-5 as the distribution was the same as the GM-10. However, all LFE units were statistically significant between the distribution of soil associations.

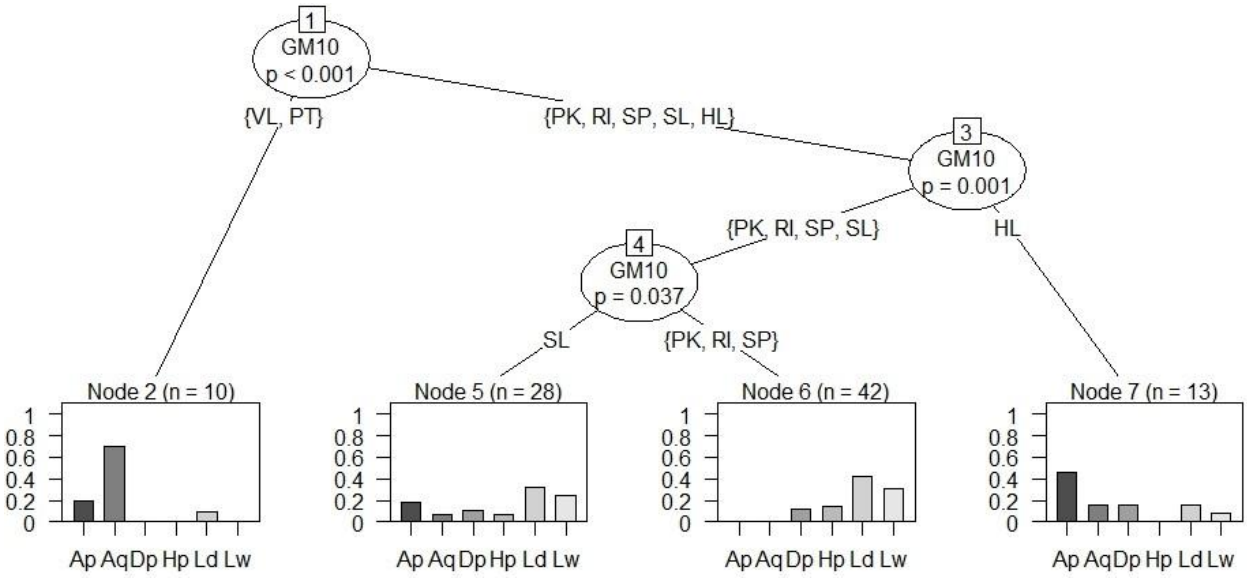


Figure 3.6: GM-10 decision tree splits on apedal (Ap), aquic (Aq), duplex (Dp), hard plinthite (Hp), lithic dry (Ld) and lithic wet (Lw) ($p < 0.05$). Abbreviation for landform elements classified are: PK – peak, RI – ridge, SP – spur, SL – slope, HL – hollow, VL -valley, PT – pit.

The GM-5 is shown in Figure 3.7. As with the GM-10, not all LFEs were classified on this area as a scarp is not present and therefore, the GM-5 only has four TMU units. Unlike the GM-10 and expert LFE, the GM-5 classified crests with a higher percent area (39%) than slopes (34%), while foot-slopes occupied 15% and valleys occupied 12% of the area. Therefore, there is a gradient from crest to valley positions in terms of percent area of each LFE. This was attributed to the aggregation method, as the aggregation was based on soil homogeneity and not landscape similarity. For example, spurs are

sloping microfeatures however, spurs were not grouped with slopes because they showed a difference in the distribution of soil associations. Instead, spurs were grouped with peaks and ridges creating the large area of crest positions.

To determine the difference between GM-5 and the expert LFE, the expert system was overlaid on the GM-5. At each pixel, it was determined if the values were the same. If the classification was the same, the pixel was marked as a match. The difference between the two systems is shown in Figure 3.7. The two systems disagreed in 57% of the area. As expected, the most agreement was seen on crest and valley positions and the most disagreement on mid-slopes and foot-slopes. This was expected because in general, crests and valley positions are fairly easily to identify visually. On the other hand, the transition between mid-slopes and foot-slopes is more difficult to visually distinguish. Therefore, the areas which do not agree can be seen as areas of uncertainty between the GM-5 and expert interpretation and thus, the boundaries of the LFEs.

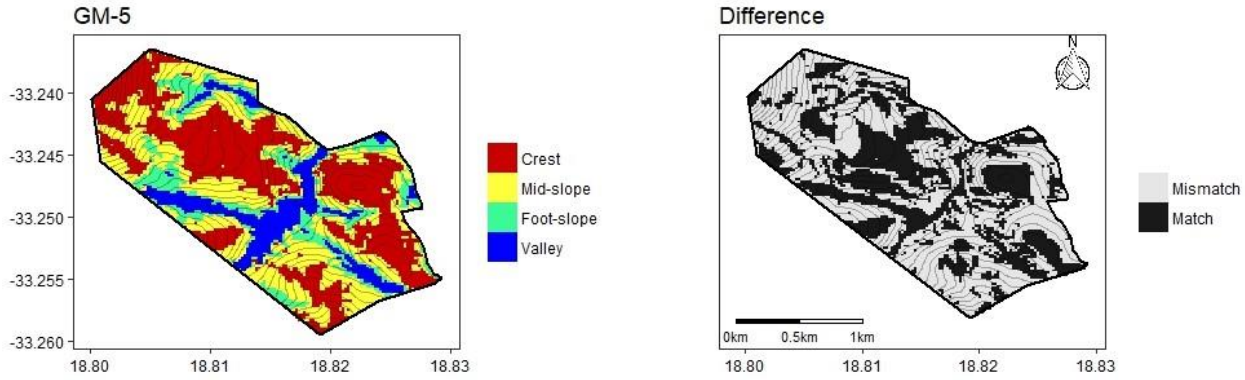


Figure 3.7: GM-5 and the difference between the GM-5 and expert LFE overlaid with 5 m contour lines.

3.3.3 Soil association distribution

No one soil association was incorporated into one LFE unit. This was attributed to the complex nature of the site where soils were observed where they are not expected. For example, Ld soils such as Lithic Halpoxerepts/Haploxeralfs are expected primarily on convex landscape positions due to erosional processes exceeding pedogenesis (Scholms et al., 1983). Although most Ld soils occur on higher elevation slopes and crest positions, some are found on foot-slopes and valleys where soil moisture is expected.

The wet variants of these lithic soils were found, surprisingly, on crest and mid-slope positions. The wetness of the Lw soils was found either as an albic horizon or as mottles in the cambic horizon. These signs of wetness can be contributed to the rolling topography with broad summits. Therefore, water can move as subsurface water flow on crests and mid-slopes and accumulate in lower positions. Additionally, this suggests that there are other environmental factors contributing to the spatial distribution of Lw and Ld soils.

The Dp soils were found on sloping positions where lighter textured, often gravelly, topsoil/albic horizons overlie a strongly structured, clay rich subsoil. The strongly structured subsoil is usually a result of in-situ weathering of the shale parent material, while the lighter textured, gravelly topsoil material is derived from local colluvial creep (Scholms et al., 1983). These soils represent a more advanced stage of pedogenic stability as sufficient time has evolved to weather and transform the shale into an aggregated soil material.

The Hp soils were found on mid-elevation crest positions where they are typically not expected. These soils are mostly comprised of moderately hard relic plinthite which can be seen as red concretions in Aeric Plinthoaquoxs and as thin bands in Plinthic Petraquepts at similar elevations. These soils were formed within pre-weathered drift material originating from the “African Surface”. In the Western Cape, red plinthite is relic because the current climate is not capable of producing hematite plinthite (le Roux and du Preez, 2006; Tyson, 1986). Therefore, the plinthite formed in a more tropical climate and the petroplinthite layer was exposed by erosion (Southard and Buol, 1988). There were also signs of plinthite degradation and re-cementation of goethite suggesting these Hp soils still have a water table present. This goethite is more stable, resistant to erosion, and could be the cause of the landscape inversion (Lambrechts, 1983). It is hypothesised that these soils are controlling the geomorphology of the region and therefore, geomorphon classification.

The Ap distribution is a little less clear as these soils were only found in small intermittent valleys throughout the field. However, the GM-5 classified Ap soils on all units. In the field, it was determined that the Ap soils have a different texture and sand grade from the other soils. Therefore, it is hypothesised that these soils developed from different depositional processes such as colluvial wash or aeolian sands (Lambrechts, 1983). These soils also represent the highest productive potential with no gravel, loamy texture, mostly friable and non-hard setting, and are physically deep.

The Aq soils were stratified well by the GM-5 where they are found in valleys and foot-slopes. This soil association is comprised of Typic Endoaqualfs/Epiaquepts and Plinthic Haplaquoxs. The Plinthic Haplaquoxs were found on foot-slope positions and have an albic horizon with discrete Fe-Mn nodules located throughout. These soils developed just above a seep where a fluctuating water table results in the accumulation of Fe and Mn, that has been mobilized from the "African surface" (Scholms et al., 1983). The other Aq soils are located near streams and erosion gullies causing a shallow water table.

3.3.4 Soil property stratification

The GM-5 was used to display how well geomorphons stratify soil properties because the GM-5 had more homogenous residuals as the other two LFE systems did not have enough samples per unit. For example, on the GM-10, there was only four samples on peak and one sample on pit positions. In contrast, there were 41 samples on crest and 10 samples on valley positions for the GM-5. A REML was selected because it does not assume a randomised design as the systemic sampling design is not random. Additionally, REML be implemented on samples that are correlated in space and time (Lark and Cullis, 2004). The REML model diagnostics and parameters for the GM-5 are shown in Table 3.2. Besides L colour, all soil properties had a variance function indicating that the mean is dependent on the variance which changes over each unit (Lark et al., 2000). Therefore, the GM-5 units are not capturing all of the variability of the soil properties. It should be noted that L colour, SOC, ERD, sand, and silt did not have homogeneous variance between groups, however, all soil properties had normal residuals

Table 3.2: REML model diagnostics and parameters showing Akaike information criterion (AIC), transformation, variance function (Variance), variogram function (Variogram), and P values for the Levene's Test (Heterogeneity) and Shapiro test (Normality).

Property	AIC	Variance	Variogram	Heterogeneity	Normality
L colour	-338	-	Spherical	0.50	<0.05
EC	-68.5	Power	Linear	<0.01	<0.01
SOC	120	Power	Exponential	0.08	<0.01
ERD	865	Power	Exponential	0.18	<0.01
DL	909	Power	Spherical	<0.05	<0.01
Gravel	719	Exponential	Spherical	<0.01	<0.01
Sand	638	Exponential	Exponential	0.14	<0.01
Silt	593	Exponential	Exponential	0.20	<0.01
Clay	480	Power	Exponential	<0.01	<0.05

The Tukey-Kramer comparison on each GM-5 unit is shown in Figure 3.8. All soil properties showed statistically significant differences of means between at least two GM-5 units except soil EC. The log of soil L colour showed a significant increase in valleys. These planar landscape positions have resulted in wet topsoils causing bleaching through iron removal and exposing coarser silicate particles (le Roux et al., 2015). The bleaching has resulted in an increase in albedo and therefore an increased L value (Fontes, 1996).

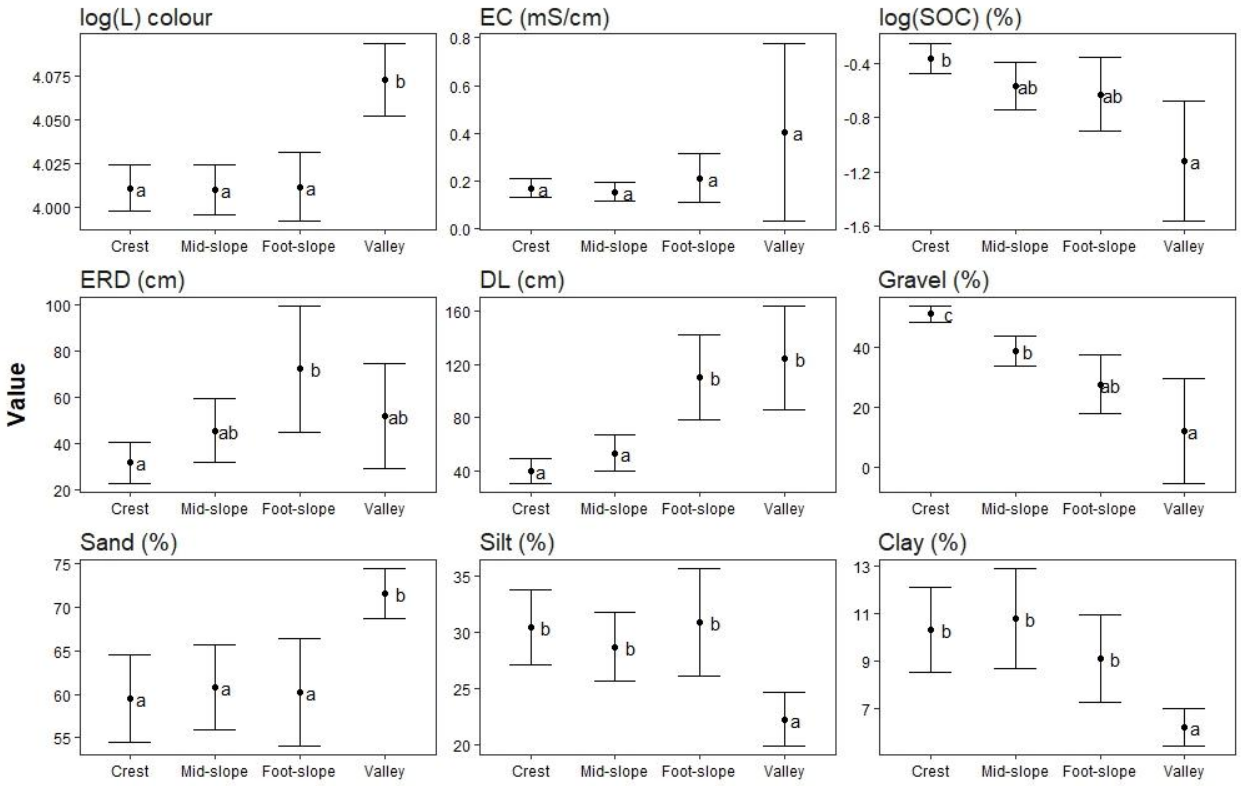


Figure 3.8: Shows the mean soil property with 95% confidence intervals from stratifying the soil-landscape with the GM-5 according to the REML analysis and a post hoc Tukey-Kramer test ($p<0.05$).

Soil organic carbon showed a decreasing trend from crest to valley positions. This was unexpected because higher SOC is anticipated on places were moisture accumulates (Safadoust et al., 2015). This is partially explained by the higher clay content on crests and mid-slopes which can stabilise SOC through sorption and micro-aggregation (Singh et al., 2017). Additionally, land use management practices such as leaving the field fallow can inhibit SOC accumulation (Dilling and Failey, 2013). This may explain the low SOC on valley positions, which due to their flat surfaces and management

practices, are more susceptible to loss of micro-aggregation through physical displacement (Hillel, 1998).

Effective rooting depth showed a difference between crest and foot-slope positions while DL showed an increasing trend down slope. These results were expected as ERD is not only a function of depth to impermeable layers but also wetness. Thus, there was a large variation throughout the landscape with crest and mid-slope positions having lithic contact (some plinthic contact), foot-slope positions showing particle accumulation, and valley positions having a shallow water table. The clear trend in DL is a result of depositional processes on lower elevations and also plinthic contact at mid-elevation positions. However, these results are in contrast to the findings of Flynn et al. (2019), as when disaggregating the LTS (through geomorphons), showed a clear trend in ERD classes from crest to valley positions.

Gravel was separated well through the GM-5 and showed a decreasing trend downslope with signs of preferential erosion. Preferential erosion has removed the finer particles on higher elevation positions and deposited them on foot-slope and valley positions (Shi and Schulin, 2018). This has resulted in the residual accumulation of the larger gravel particles upslope.

Sand, silt, and clay results did not show clear signs of preferential erosion. Sand had a weak increasing trend downslope while silt and clay showed the opposite trend. The weak trends were also observed by Silva et al. (2016), who determined that the highly weathered surface has led to homogenise particle size distribution. The increase in sand downslope may be due to fluvial sands resulting in an absolute increase in sand content on lower surfaces. Silt and clay trends are best explained by the parent material found on the site. Soils at higher elevations developed from residual, highly weathered material from the old African surface (Lambrechts, 1983; Scholms et al., 1983). The soils developed from this parent material, have a higher Fe content preventing their removal by stabilising the soil aggregates (Barral et al., 1998).

These results indicate that the aggregated GM-5 corresponds well to slope shape which affects the soil distribution at this scale. For example, crest represent erosion zones, mid-slope represents areas of transportation, and valleys represent areas of accumulation. This is apparent from the soil

properties found on crest (high gravel) and valley positions (high sand) and the GM-5 makes pedological sense.

3.3.5 Soil property predictions

Soil properties were predicted through ordinary least squares with bootstrap resamples. The bootstrap resampling was performed with 2000 iterations because 2000 bootstraps are required to obtain a confidence interval of 99% according to Davidson and Mackinnon (2007). The bootstrapped RMSE values for all soil properties are shown in Figure 3.9. The GM-5 was statistically the best predictor (lowest RMSE) for L colour, DL, gravel, sand, silt and clay. The GM-10 did not predict any soil properties best according to RSME values. The expert LFE was statistically the top predictor for EC, SOC, and ERD. The GM-5 accounted for a low of 10% (EC) to a high of 43% (gravel) of the soil property variance (R^2). Surprisingly, the expert LFE had a lower R^2 for EC of 3% and had a high of 24% for gravel. This suggests that the GM-5 can give an indication of the spatial distribution of many soil properties better than the other systems at this scale.

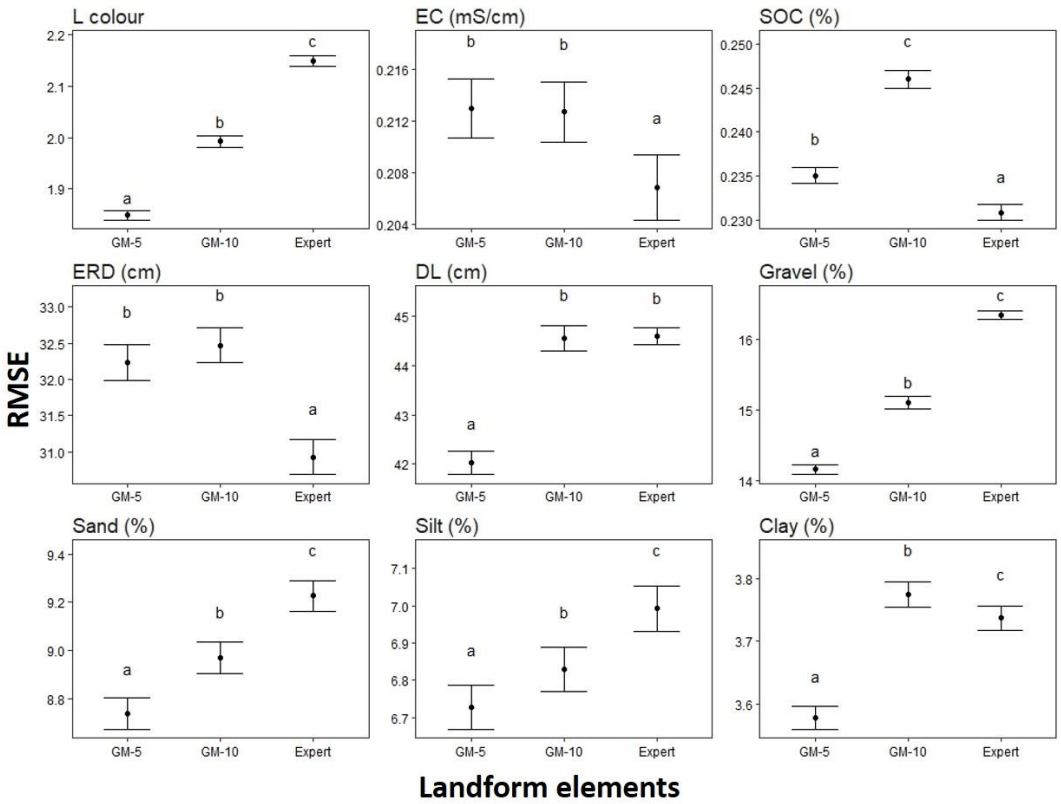


Figure 3.9: Mean RMSE values shown with 95% confidence intervals from bootstrap analysis comparing the GM-5, GM-10 and the expert LFEs (Expert) ($p<0.05$).

It is surprising that aggregating geomorphons into the GM-5 improved the results for almost all soil properties relative to the GM-10. One explanation is that there were more soil samples per unit and more balanced data for the GM-5 compared with the GM-10. The number of samples per GM-10 and GM-5 is shown in Table 3.3. Therefore, the mean within each unit was more reliable, the standard error decreased, and the dataset was more balanced for the GM-5. However, a higher density of soil samples may change these results although this would also increase the financial cost. When compared with the expert LFE with the GM-5, the GM-5 predicted more soil properties (6 vs. 3 soil properties) with a lower RMSE, however, it also shows that the best fitting LFE system will depend on the soil property of interest. For example, when predicting SOC, it would be more appropriate to use the expert LFE than the GM-5.

Table 3.3: Number of soil observations per geomorphon unit compared with the number of terrain morphological units (TMUs) number of observations.

Geomorphon units	Observations	TMUs	Observations
Peak	4	Crest	42
Ridge	8		
Spur	30		
Slope	28	Mid-slope	28
Hollow	13	Foot-slope	13
Foot-slope	0		
Valley	9	Valley	10
Depression	1		

These results correspond to the findings of Moravej et al. (2012), who found that automated landform classification can produce a more detailed and accurate map than manual delineation. The results also correspond to Libohova et al. (2016), who found 5-units could be predicted with a higher accuracy from soil morphological properties. Nevertheless, this is in contrast to Barka et al. (2011), who found that manual delineation was more accurate when comparing several automated techniques to predict forest type. However, the authors relate this to the sample design implemented.

3.4 Conclusion

Twenty-four geomorphons were derived and the best fitting geomorphon was used to stratify the soil-landscape at the farm-scale. Geomorphon selection was based on a bootstrapped CV to determine the geomorphon with the highest association with soil classes. The best fitting geomorphon was aggregated into a 5-unit system to correspond with TMUs found in the LTS. The aggregation was based on similarities of soil classes between geomorphon units using a decision tree. The aggregated geomorphon was used to see how well the system stratifies 9 soil properties. The GM-5 and GM-10 prediction potential was compared against an expert delineated system by predicting the soil properties through bootstrap analysis.

The main findings of this study are: i) The geomorphon with a 10 m resolution and 200 cell search radii had the highest association with soil classes; ii) within each GM-5 unit, the distribution of soil associations was statistically significant; iii) the GM-5 agreed with the expert LFE in 43% of the area and thus, there is a high degree of uncertainty between the GM-5 and expert LFE; iv) the GM-5 showed trends in the mean for all soil properties except EC however, some trends such as particle size distribution, only showed a weak trend; v) the GM-5 had the highest prediction potential when estimating multiple soil properties; therefore, the GM-5 can be used as an initial indication of the soil spatial distribution.

3.5 References

Atkinson, J., Rozanov, A.B., de Clercq, W.P., 2017. Evaluating the effects of generalisation approaches and DEM resolution on the extraction of terrain indices in KwaZulu Natal, South Africa. South African J. Geomatics 6, 245–261. https://doi.org/10.4314/sajg.v6i2.9

Barka, I., Vladovic, J., Frantisek, M., 2011. Landform Classification and its Application in Predictive Mapping of Soil and Forest Units. GIS Ostrava 1, 23–26.

Barral, M.T., Arias, M., Guerif, J., 1998. Effects of iron and organic matter on the porosity and structural stability of soil aggregates. Soil Tillage Res. 46, 261–272.

Bleeker, P., Speight, J.G., 1978. Soil-landform relsationships at two localities in Papua New Guinea. Geoderma 21, 183–198.

Botha, C.C., 2016. Disaggregating of land type data to aquire functional soil information (MSc thesis). University of the Free State.

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

Cressie, N., 1985. Fitting variogram models by weighted least squares. J. Int. Assoc. Math. Geol. 17, 563–586. https://doi.org/10.1007/BF01032109

Davidson, R., Mackinnon, J.G., 2007. Bootstrap tests: how many bootstraps? Econom. Rev. 19, 55–68. https://doi.org/10.1080/07474930008800459

Di Stefano, M., Mayer, L.A., 2018. An automatic procedure for the quantitative characterization of submarine bedforms. Geosciences 8, 1–28.

Dilling, L., Failey, E., 2013. Managing carbon in a multiple use world: The implications of land-use decision context for carbon management. Glob. Environ. Chang. 23, 291–300. https://doi.org/10.1016/j.gloenvcha.2012.10.012

Drăgut, L., Eisank, C., 2012. Automated object-based classification of topography from SRTM data. Geomorphology 141–142, 21–33. https://doi.org/10.1016/j.geomorph.2011.12.001

Efron, B., Tibshirani, R., 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy Authors. Stat. Sci. 1, 54–75.

Evans, I.S., 2012. Geomorphometry and landform mapping: What is a landform? Geomorphology 137, 94–106. https://doi.org/10.1016/j.geomorph.2010.09.029

FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012. Harmonized World Soil Database (version 1.2). Rome, Italy.

Flynn, T., Rozanov, A., de Clercq, W., Warr, B., Clarke, C., 2019. Semi-automatic disaggregation of a national resource inventory into a farm-scale soil depth class map. Geoderma 337, 1136–1145. https://doi.org/10.1016/j.geoderma.2018.11.003

Fontes, A.F., 1996. Soil Albedo in Relation to Soil Color, Moisture and Roughness (PhD thesis). The University of Arizona.

GRASS Development Team, 2017. Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Found.

Guillocheau, F., Simon, B., Baby, G., Bessin, P., Robin, C., Dauteuil, O., 2018. Planation surfaces as a record of mantle dynamics: The case example of Africa. Gondwana Res. 53, 82–98. https://doi.org/10.1016/j.gr.2017.05.015

Hillel, D., 1998. Environmental Soil Physics, Journal of Environment Quality. Academic Press, San Diego, CA. https://doi.org/10.2134/jeq1999.00472425002800060046x

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. J. Comput. Graph. Stat. 15, 651–674.

Huggett, R.J., 1975. Soil landscape systems: A model of soil genesis. Geoderma 13, 1–22.

Irvin, B.J., Ventura, S.J., Slater, B.K., 1997. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. Geoderma 77, 137–154. https://doi.org/10.1016/S0016-7061(97)00019-0

Iwahashi, J., Pike, R.J., 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology 86, 409–440. https://doi.org/10.1016/j.geomorph.2006.09.012

Jasiewicz, J., Netzel, P., Stepinski, T.F., 2014. Geomorphology Landscape similarity, retrieval, and machine mapping of physiographic units. Geomorphology 221, 104–112. https://doi.org/10.1016/j.geomorph.2014.06.011

Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons - a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156. https://doi.org/10.1016/j.geomorph.2012.11.005

Jenness, J., 2006. Topographic Position Index (TPI). Jenness Enterp.

Kramm, T., Hoffmeister, D., Curdt, C., Maleki, S., Khormali, F., Kehl, M., 2017. Accuracy Assessment of Landform

Classification Approaches on Different Spatial Scales for the Iranian Loess Plateau. Int. J. Geo-Information 6, 1–22. https://doi.org/10.3390/ijgi6110366

Lambrechts, J.J.N., 1983. Soils, Soil Process and Distribution in the Fynbos Region: An Introduction. Council for Scientific and Industral Research, Pretoria.

Land Type Survey Staff, 1972-2006. Land Types of South Africa on 1:250 000 scale. Pretoria, South Africa.

Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. Eur. J. Soil Sci. 55, 799–813. https://doi.org/10.1111/j.1365-2389.2004.00637.x

Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 97, 787–799. https://doi.org/10.1111/j.1365-2389.2005.00768.x

le Roux, J.L., de Clercq, W.P., Clarke, C., 2015. The occurence of bleached topsoils on weakly structured subsoil horizons in the Western Cape and Mpumalanga provinces of South Africa (MSc thesis). Stellenbosch University.

Libohova, Z., Winzeler, H.E., Lee, B., Schoeneberger, P.J., Datta, J., Owens, P.R., 2016. Geomorphons: Landform and property predictions in a glacial moraine in Indiana landscapes. Catena 142, 66–76. https://doi.org/10.1016/j.catena.2016.01.002

Liebetrau, M.A., 1983. Measures of association, in: Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park, California, pp. 15–16.

MacMillan, R.A., Shary, P.A., 2009. Landforms and landform elements in geomorphometry, in: Developments in Soil Science. Elsevier B.V., pp. 227–254. https://doi.org/10.1016/S0166-2481(08)00009-3

McBratney, A.B., Santos, M.L.M., Minasnya, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142, 285–293. https://doi.org/10.1016/j.geoderma.2007.08.022

Moore, A.W., Russell, J.S., 1966. Potential use of numerical analysis and Adansonian concepts in soil science. Aust. J. Sci. 29, 141–142.

Moravej, K., Eghbal, M.K., Toomanian, N., Mahmoodi, S., 2012. Comparison of automated and manual landform delineation in semi detailed soil survey procedure. African J. Agric. Res. 7, 2592–2600. https://doi.org/10.5897/AJAR11.728

Odeh, I.O.A., Mcbratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197–214.

Partridge, T.C., Maud, R.R., 1987. The Geomorphic Evolution of Southern Africa Since the Mesozoic. South African J. Geol. 90, 179–208.

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

Pinheiro, J.C., Bates, D.M., DebRoy, S., Sarkar, D., R Core Team, 2018. nlme: Linear and Nonlinear Mixed Effects Model. R Packag. version 3.1-131.1.

Pinto, L.C., de Mello, C.R., Norton, L.D., Owens, P.R., Curi, N., 2016. Spatial prediction of soil-water transmissivity based on fuzzy logic in a Brazilian headwater watershed. Catena 143, 26–34. https://doi.org/10.1016/j.catena.2016.03.033

Roecker, S.M., Thompson, J.A., 2010. Scale Effects on Terrain Attribute Calculation and Their Use as Environmental Covariates for Digital Soil Mapping, in: Boettinger, J.L., Howell, D.W., Moore, A.C.,

Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping. Springer, Dordrecht, pp. 55–66. https://doi.org/10.1007/978-90-481-8863-5

Roux, P.A.L., Preez, C.C., 2006. Nature and distribution of South African plinthic soils: Conditions for occurrence of soft and hard plinthic soils. South African J. Plant Soil 23, 120–125. https://doi.org/10.1080/02571862.2006.10634741

Ruhe, R. V., 1960. Elements of soil landscape, in: 7th International Congress of Soil Science. pp. 165–170.

Safadoust, A., Doaei, N., Mahboubi, A.A., Mosaddeghi, M.R., 2015. Arid Land Research and Management Long-term Cultivation and Landscape Position Effects on Aggregate Size and Organic Carbon Fractionation on Surface Soil Properties in Semi-arid Region of Iran, in: Arid Land Research and Management. Taylor & Francis, pp. 1–18. https://doi.org/10.1080/15324982.2015.1016244

Schmidt, J., Hewitt, A., 2004. Fuzzy land element classification from DTMs based on geometry and terrain position. Geoderma 121, 243–256. https://doi.org/10.1016/j.geoderma.2003.10.008

Scholms, B.H.A., Ellis, F., Lambrechts, J.J.N., 1983. Soils of the Cape Coastal Platform, in: Deacon, H.J., Hendey, Q.B., Lambrechts, J.J.N. (Eds.), . Council for Scientific and Industral Research, Pretoria.

Shi, P., Schulin, R., 2018. Erosion-induced losses of carbon, nitrogen, phosphorus and heavy metals from agricultural soils of contrasting organic matter management. Sci. Total Environ. 618, 210–218. https://doi.org/10.1016/j.scitotenv.2017.11.060

Silva, S.H.G., Menezes, M.D., Mello, C.R., Góes, H.T.P., Owens, P.R., Curi, N., 2016. Geomorphometric tool associated with soil types and properties spatial variability at watersheds under tropical conditions. Sci. Agric. 73, 363–370. https://doi.org/10.1590/0103-9016-2015-0293

Singh, M., Sarkar, B., Sarkar, S., Churchman, J., Bolan, N., Mandal, S., Menon, M., Purakayastha, T.J., Beerling, D.J., 2017. Stabilization of Soil Organic Carbon as Influenced by Clay Mineralogy Stabilization of Soil Organic Carbon as Influenced by Clay Mineralogy. Adv. Agron. 148, 1–51. https://doi.org/10.1016/bs.agron.2017.11.001

Soil Classification Working Group, 1991. Soil Classificatioin: a Taxonomic System for South Africa, 2nd ed. Department of Agricultural Development, Pretoria, South Africa.

Soil Survey Staff, 2017. Soil Survey Manual Agriculture. Handbook 18. USDA, Nat. Resour. Conserv. Serv. 18, 483. https://doi.org/10.1097/00010694-195112000-00022

Soil Survey Staff, 2014. Keys to soil taxonomy, 12th ed. USDA-Natural Resources Conservation Service, Washington, DC.

Southard, R., Buol, S., 1988. Subsoil Saturated Hydraulic Conductivity in Relation to Soil Properties in the North Carolina Coastal Plain. Soil Sci. Soc. Am. 52, 1091–1094.

Sufahani, S.F., Ahmad, A., 2012. A Comparison between Normal and Non- Normal Data in Bootstrap. Appl. Math. Sci. 6, 4547–4560.

Tobler, W.R., 1987. Measuring Spatial Resolution, in: Land Resources Information System. Beijinig, pp. 12–16.

Tonkin, P.J., Basher, L.R., 1990. Soil-stratigraphic techniques in the study of soil and landform evolution across the Southern Alps, New Zealand. Geomorphology 3, 547–575. https://doi.org/10.1016/0169-555X(90)90020-Q

Tyson, P.D., 1986. Climate change and variability in southern Africa. Oxford Unviersity Press, Cape Town. https://doi.org/doi.org/10.1002/joc.3370080411

van Zijl, G., van Tol, J., Tinnefeld, M., Le Roux, P., 2019. A hillslope based digital soil mapping approach, for hydropedological assessments. Geoderma 354, 113888. https://doi.org/10.1016/j.geoderma.2019.113888

Weiss, A.D., 2000. Topographic Position and Landforms Analysis, in: ESRI User Conference.

Zhang, W., Montgomery, D.R., 1994. Digital elevation model grid size, landscape representation, and hydrological simulation. Water Resour. Res. 30, 1019–1028.

Zhong, T., Cang, X., Ruoyin, L., Tang, G., 2009. Landform Classification Based on Hillslope Units from DEMs, in: 30th Asian Conference on Remote Sensing 2009.

Chapter 4 Semi-automatic disaggregation of a national resource inventory into a farm-scale soil depth class map

This chapter is based on the publication Flynn, T., Rozanov, A., de Clercq, W., Warr, B., Clarke, C. 2019. Semi-automatic disaggregation of a national resource inventory into a farm-scale soil depth class map. *Geoderma* 337, 1136-1145 found in Appendix B.

Abstract:

Knowledge of soil depth spatial variability is important for land use management especially in dryland agriculture regions, which rely on climate and soils to provide adequate water and nutrients during the growing season. Soil spatial variability can be predicted from legacy soil data through machine learning techniques producing quantitative soil maps requiring minimal resources. South Africa has a country wide 1:250,000 scale resource map known as the Land Type Survey (LTS) which includes soil properties such as soil depth, soil class, root limiting layer, clay content, and texture. Each LTS polygon (land type), is comprised of unique soil – terrain patterns and is therefore, not a true soil map. This study aims to disaggregate the LTS into a farm-scale soil depth class map through a two-step disaggregation approach. First, landform elements (LFEs) were predicted through a pattern recognition algorithm known as geomorphons. Geomorphons, together with the original LTS were overlaid to produce polygons with unique distributions of soil. The polygons were disaggregated further to produce a raster map of soil depth classes through a soil map disaggregation algorithm known as DSMART. The first most probable class raster achieved an accuracy of 68% and for the two most probable class rasters, an accuracy of 90% was achieved. The two-step approach proved necessary for producing a farm-scale soil map. The result of this study is significant as it produced a soil depth class map from a national resource map at a scale and resolution (10 m) suitable for farm management.

4.1 Introduction

Soil depth is one of the seven major considerations when evaluating soil quality (Bunning et al., 2011). Soil depth and its spatial variability is crucially important for activities such as planning irrigation (Myburgh et al., 1996), hydrological modelling (Devia et al., 2015), and estimating soil carbon stocks (Wiese et al., 2016). Soil properties, including depth, may be highly variable on old land surfaces due to multiple cycles of erosion and deposition (Rozanov et al., 2017). Soil depth is difficult to estimate because subsurface properties might not be detected through ancillary data (e.g., unexposed hard

plinthite), and usually requires physical observations which are expensive. Alternatively, soil depth can be predicted through DSM which requires fewer soil samples and, when soil legacy data is available, no soil samples at all. As increasingly high-resolution ancillary data becomes available, DSM techniques such as the *scorpan* framework (McBratney et al., 2003), can be adapted to and make better use of legacy data, for instance, by generating farm-scale (1:20,000) maps at an appropriate resolution.

The LTS (Land Type Survey Staff, 1972 - 2006) was developed in a hierarchical structure that is spatially scaled from the top down. There are many difficulties facing the disaggregation of the LTS, such as the coarse spatial extent (>160 ha), coarse soil class scale, and the fact that the LTS consists of complex soil-terrain polygons. To generate finer-scale soil information, TMUs can be classified to disaggregate the coarse land type polygons into basic TMUs that have associated soil attributes. In doing so, the LTS can be stratified into polygons with more homogenous distribution of soil. Therefore, each stratified polygon has a unique distribution of soil consisting of unique patterns of the spatial soil distribution. This strategy forms a toposequence in each polygon for each micro-climate zone, however, the toposequence and the soil distribution are estimated as percent area within each land type. Therefore, their specific geographic distribution is unknown.

There are many LFE classification algorithms which include nested-clustering methods (Iwahashi and Pike, 2007), fuzzy landform classification (Irvin et al., 1997; Schmidt and Hewitt, 2004), object-based classification (Dragut, 2011), and more. However, these algorithms are not flexible to changes in scale, are often computationally inefficient, and/or do not coincide with slope position used by soil scientists in delineating the original TMUs (Jasiewicz and Stepinski, 2013; Libohova et al., 2016).

Developed by Jasiewicz and Stepinski (2013), geomorphons is a pattern recognition algorithm created to be computationally efficient, flexible to scale through an automatically adjusting moving window, and when geomorphons units are aggregated, can correspond to slope position (Libohova et al., 2016). Geomorphons classify the ten most common LFEs known as flat (FL), peak (PK), ridge (RI), shoulder (SH), spur (SP), slope (SL), hollow (HL), foot-slope (FS), valley (VL), and pit (PT). Silva et al. (2016) demonstrated the flexibility of the geomorphon approach by stratifying a tropical soil-landscape to help soil surveying in two different catchments in Brazil. Libohova et al. (2016) showed that by aggregating 10-unit geomorphon (GM-10) into a 5-unit geomorphon (GM-5) that

correspond to slope position such as foot-slope and toe-slope, geomorphons could be predicted with an 81% accuracy from soil morphological properties.

Other factors making the disaggregation of the LTS difficult is that there are very few georeferenced soil profiles and many land types do not have any georeferenced soil profiles. The lack of soil profiles makes updating the LTS's detail problematic without an additional soil survey. Additionally, the LTS was developed by different soil scientists who surveyed different land types and the understanding of how to delineate the land types developed through its creation. Therefore, different soil environmental rules were established depending on who the surveyor was and the time of the survey.

There are many approaches that have been developed that incorporate existing resource inventories into the DSM framework (Grunwald, 2006; Minasny and McBratney, 2016; Scull et al., 2003). These techniques include geostatistics, expert knowledge systems, and machine learning algorithms which have been successfully applied through kriging with external drift (Kempen et al., 2015), fuzzy logic (MacMillan et al., 2000; Silva et al., 2014; Smith et al., 2010; Yang et al., 2011), k-means clustering (Bui and Moran, 2001), decision trees (Nauman and Thompson, 2014; Sarmento et al., 2017; Subburayalu et al., 2014), and random forest (Häring et al., 2012; Nauman et al., 2014). An expert knowledge system through a fuzzy logic inference system known as SoLIM (Zhu, 1997), has been successfully applied to update the LTS into soil associations by van Zijl et al. (2013). However, these techniques often require additional soil samples, are restricted to soil polygon boundaries, and/or predict individual soil classes separately.

DSMART developed by Odgers et al. (2014), shows promise to disaggregate the LTS into a soil map. The DSMART algorithm uses resampled classification trees to create multiple realisations from soil legacy polygons. For each realisation, DSMART finds soil environmental relationships through the randomly assigned samples and covariates. These realisations are used to calculate the probability of each soil class and make the final predictions of a specified number of probable class rasters. A benefit of the DSMART algorithm is that it can predict soil classes across soil polygon boundaries and predict all soil classes simultaneously. This would be beneficial to disaggregate the LTS where soil information needs to be predicted across boundaries created by different surveyors and TMUs. Additionally, DSMART does not necessarily need to be implemented with classification trees, making it flexible to algorithm chosen and soil attribute to be predicted.

Odgers et al. (2014) who developed DSMART, showed DSMART's potential by disaggregating a soil legacy map in central Queensland, Australia. The authors achieved a 22.5% accuracy for the first most probable class raster. However, the three most probable class rasters combined, achieved an accuracy of 50%. Vincent et al. (2018) implemented DSMART to disaggregate a French resource map at a 1:250,000 scale. The authors achieved an accuracy of 41% to 72% depending on the validation technique used. Holmes et al. (2015) disaggregated a soil – terrain polygon map through DSMART to predict Soil Groups of Western Australia. The authors achieved an accuracy of 41% for the three most probable class rasters. Perhaps the most remarkable implementation of DSMART came from Chaney et al. (2016) who implemented DSMART with the Random Forest algorithm to disaggregate the Soil Survey Geographic database for the whole contiguous United States. The result of this study was a continuous soil series database known as POLARIS, which according to the first ten realisations, matched 55% of the Soil Survey Geographic database. These studies all focused on regional or even country scale and the algorithm has not yet been tested for farm-scale mapping (Malone et al., 2017).

This objective (Framework 2, Objective 3) of this study was to disaggregate the LTS into a farm-scale soil map at a resolution suitable for farming. Due to the importance of soil depth, this is the property selected for mapping. The main idea behind this implementation was to adapt a large spatial scale DSM framework and adjust it for the local needs in South Africa as many farmers do not have access to soil information. This can potentially increase the accessibility and usefulness of the LTS information to farmers by quantifying the spatial variability within the LTS and producing soil maps in a cost-effective manner

4.2 Methods and materials

4.2.1 Site description

The site location is shown and described in Chapter 2 and 3. The predominant land types in the Sandspruit catchment are shown in Figure 4.1. A small area was established to determine if the LTS can be disaggregated and downscaled into a larger scale map from the original 1:250,000 LTS polygons. The site was also selected due to the multiple land types that intersect the area.

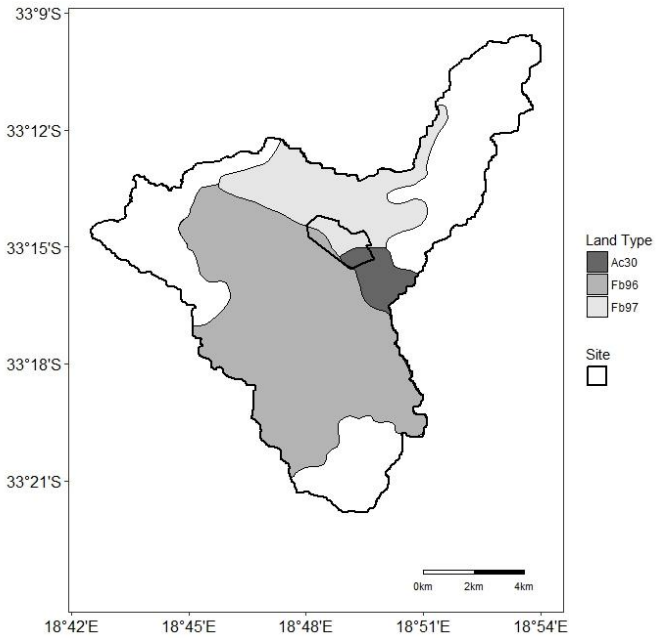


Figure 4.1: Showing the dominant land types in the Sandspruit catchment and the research site within the catchment.

There are three land types that dissect the site according to the Land Type Survey Staff, (1972-2006). These land types are Ac30, Fb96, and Fb97 as displayed on the land type sheet 3318 (Cape Town). There are two dominant soils in the catchment according the LTS. These are the relatively shallow residual soils such as Lithic Haploxerepts and the deeper red apedal soils such as Typic Haploxerept. These soils represent different parent materials and different ages. The residual soils are weathered from the shale parent material and are overlain by a thin, loamy, usually gravely creep layer. These residual soils are younger and are found at lower altitudes than the highly weathered drift from the old African surface which occurs on pre-weathered shale. The red apedal soils, form within this highly weathered drift material and are found at higher altitudes.

4.2.2 Software

Geomorphons and covariates were developed from a digital elevation model (DEM) using the Geographic Resource Analysis Support System (GRASS v7.2; GRASS Development Team, 2017) and the System for Automated Geoscientific Analyses (SAGA v2.3.2; Conrad et al., 2015), respectively. All models and statistical analysis were conducted in R software (R Core Team, 2017). Disaggregation was developed from the DSMART algorithm as in the rdsmart package (Odgers and Malone, 2017). However, the rdsmart code was adapted to incorporate any classification model available in the caret R package (Kuhn et al., 2018).

4.2.3 Land Type Survey database

The land types were obtained from the LTS sheet 3318 (Cape Town) which has a basic map unit of 160 ha. The legends for each land type were gathered from separate text files. A depiction of the structure of the LTS data is shown in Table 4.1. The 3 land types depth ranged from 0 to over 120 cm depending on the soil type. For example, on land type Ac30, Gs soils (Lithic Haploxerepts) ranged from 200 to 500 cm and Oa soils (Typic Haploxerepts) were all over 900 cm according to the LTS. Therefore, a histogram of the mean soil depth for each soil type was plotted and depth ranges were taken in an attempt to get an equal distribution of soil observations within each depth class. Based on the histogram, the soils were divided into shallow (0 cm - 40 cm), moderate (40 cm – 80 cm) and deep classes (more than 80 cm).

Table 4.1: The hierarchical structure of the data obtained for the LTS information, how the information is represented and how the files are obtained.

Information	Representation	File type
Land types	Spatial polygons	Shapefile
TMUs	Probabilities in each polygon	Legend
Soil attributes	Probabilities in each TMU	Legend

4.2.4 Digital elevation model

The resampled DEM described in Chapter 2 was used to develop a range of topographic covariates. The DEM was re-projected into the Hartebeeshoek94 datum projected coordinate system. Reprojection was necessary because the data sets were distributed on different coordinates systems, accurate distances between points was needed, and to locate the soil observations in the field. However, this can add distortion to the DEM and lower its accuracy (Hengl and Evans, 2009).

4.2.5 Disaggregation approach

The disaggregation approach in this paper follows a two-step method to disaggregate the LTS shown in Figure 4.2. The first step of the processes was to classify TMUs corresponding to the LTS legend. The TMUs were overlaid with the LTS and soil depth class probabilities (percent area specified in the LTS) were manually assigned to each TMU based on specific soil class depths. This created TMUs with unique distributions of soil depth classes. The TMU stratification procedure creates polygons with a more detailed spatial scale than the original LTS. DSMART was run by drawing random samples from the TMUs for 100 realisations. The covariate values were then extracted to establish soil-

environmental relationships to predict soil depth classes. The final product consisted of the first and second most probable class rasters, the probabilities of each soil depth class, and spatial uncertainties. Spatial uncertainties are an evaluation of how certain the model is of the predictions made through the extent of the area. This process was compared to resample polygons created from a manually delineated TMUs and using the LTS polygons with no TMU classification.

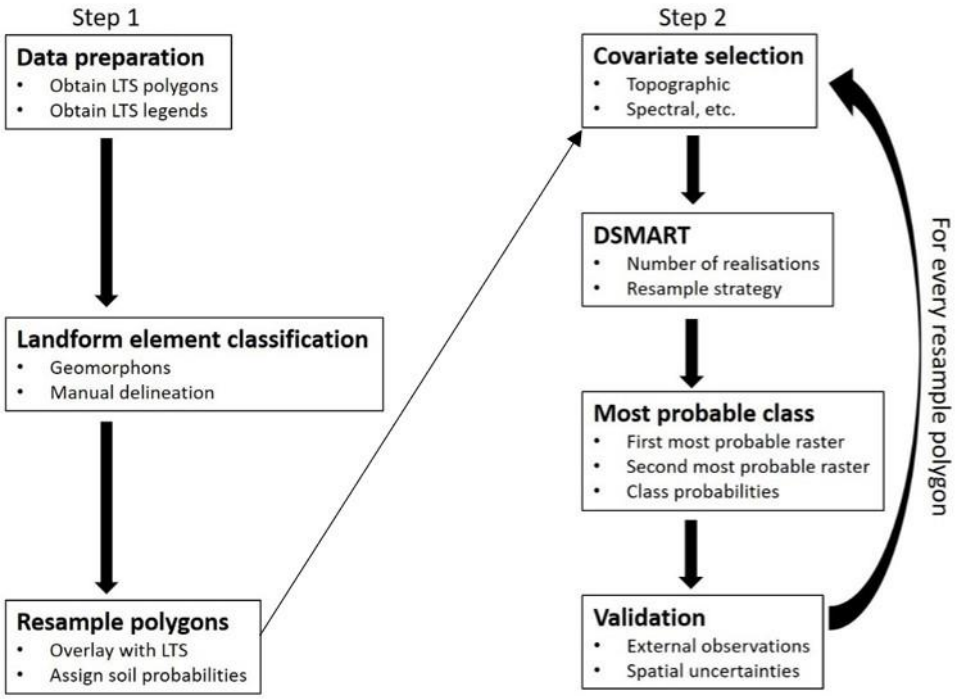


Figure 4.2: Methodology for the disaggregation of the LTS by stratifying the LTS through geomorphons and running DSMART to extract the spatial distribution of the soil depth classes.

4.2.6 Landform element development

A geomorphon was developed through the r.geomorphons GRASS GIS add-on developed by Jasiewicz and Stepinski (2013). The final geomorphon selected had a 200-cell search radius and a flatness threshold of 1 degree. It was thought that this geomorphon represented the rolling topography best and was most suitable to vectorise into polygons. To correspond with TMUs, the GM-10 (LFEs) was aggregated into a GM-5 (TMUs). It should be noted that the LTS has up to five TMUs, however, not all land types have all 5-units. For example, some land types only have foot-slopes and valleys, therefore, the aggregation procedure needs to correspond to these units to match the LTS legend.

To aggregate the geomorphon units, the geomorphon values were changed for each individual pixel of the original GM-10. Peak, ridge, and shoulder were aggregated into crest as these positions are convex, usually high elevation positions with slopes accruing below. Spur and slope are generally found on mid-elevation positions and were therefore, aggregated into mid-slope. Hollow and foot-slope positions, which correspond to concave slopes, were aggregated into foot-slopes. Flat slopes were only found in valley positions and along with valley and pit, were aggregated into valley. The reclassification method used in this study is a simplification of the method developed by Libohova et al. (2016), who aggregated GM-10 into GM-5 based on a slope gradient threshold. However, no slope gradient threshold was used for aggregation. The TMU aggregation and the original geomorphons units are shown in Table 4.2.

Table 4.2: Shows the aggregated geomorphons units (TMUs) vs. the original geomorphons units.

TMUs	Geomorphon
Crest	Peak
	Ridge
	Shoulder
Slope	Spur
	Slope
Foot-slope	Hollow
	Foot-slope
Valley	Valley
	Depression
	Flat

To determine if the geomorphon algorithm can mimic the "mental model" used to define TMUs in the LTS, an additional manually delineated TMUs was created (expert GM). The expert GM was developed by Dr Freddie Ellis. Dr Ellis was the surveyor for the LTS sheet 3318 (Cape Town) that covers the study area.

4.2.7 Terrain morphological units

Terrain morphological units were created for both the GM-5 and the expert GM by first converting the LFE rasters into polygons and then overlaying with the land types. These TMUs will be referred to as LTS-GM5 and the LTS-EX5, respectively. This can be seen as the first step in disaggregating the LTS as the TMUs disaggregate the land types from soil-terrain patterns, into soil patterns. The comparison between the two TMUs is important as it is thought that the LTS-GM5 must mimic the

TMUs generated by the LTS surveyor to obtain accurate results. This is because the soil distribution was estimated from the toposequences observed in the field.

To determine if the two-step disaggregation approach is necessary, the LTS was also disaggregated without TMU classification in a single step approach. This was done by taking the total soil depth class probabilities of each LTS legend. Therefore, in this implementation, the land types themselves were used as resample polygons. These resample polygons will be referred to as LTU. This comparison is important as the two-step disaggregation approach requires a greater in-depth knowledge of soil environmental relationships and since it requires two models, the two-step approach may have errors which propagate through each step. For example, the geomorphons might misclassify a LFE which will lead to incorrect soil depth class allocation in DSMART. However, this error propagation was not accounted for in this study.

4.2.8 Covariate development

A pool of covariates was developed that were intended to represent relief, organisms, parent material, and neighbourhood according to the *scorpan* factors (McBratney et al., 2003). The covariates developed were altitude, convexity, plan curvature, profile curvature, negative openness, SAGA wetness index (SWI), slope gradient, slope length factor (LS factor), stream power index (SPI), topographic position index (TPI), vertical distance to stream (VDS), soil adjusted vegetative index (SAVI; Huete, 1988), lithology, soil colour index (SCI), and soil redness index (SRI; Ray et al., 2004). These covariates were selected from experience of soil modelling in the area and it was thought that they capture the environmental factors that affect soil depth sufficiently, as seen in the field.

The vegetative and soil indices were developed from the Sentinel-2A satellite obtained at a 10 m resolution. The SAVI was calculated from a September 23, 2016 image to capture the vegetative growth before harvest. The soil indices (SCI and SRI) were calculated from a February 03, 2017 image during the fallow period. Together with the lithology map, the soil indices were used as an indication of soil parent material. The bands used to calculate the indices, the equations to calculate the indices, and a description of the indices used are shown in Table 4.3. For the SAVI, the vegetative factor (L) was set to 0.1 as this is suitable for most agricultural fields (Rondeaux et al., 1996).

Table 4.3: Shows the Sentinel-2A satellite bands obtained, the equations to calculate the indices, and a description of the indices.

Bands	Central Wavelength (μm)	Resolution (m)
Blue (B)	0.490	10
Green (B)	0.560	10
Red (R)	0.665	10
Near infrared (NIR)	0.842	10
Indices	**Equation**	**Property**
Colour Index	$(\mathrm{R}-\mathrm{G})/(\mathrm{R}+\mathrm{G})$	Soil colour
Redness Index	$\mathrm{R}^2/(\mathrm{B}*\mathrm{G}^2)$	Hematite
SAVI	$\frac{\mathrm{NIR}-\mathrm{R}}{\mathrm{NIR}+\mathrm{R}+\mathrm{L}}(1+\mathrm{L})$	Chlorophyll reflectance

4.2.9 Soil depth class predictions

Multinomial logistic regression with L_2 regularised (MLR) maximum log-likelihood was the algorithm applied in DSMART; however, as previously stated, it cannot be assumed that an algorithm will work best for a particular dataset (Kuhn and Johnson, 2013). Therefore, DSMART was also run with the original C5.0 algorithm (Quinlan, 1993) and the Random Forest (RF) algorithm (Breiman, 2001) as implemented in previous studies. Additionally, as soil depth classes are ordinal in nature as they are easily ranked (shallow < moderate < deep), an ordinal logistic regression (OLR) was also run. Therefore, four machine learning algorithms were implemented in DSMART. However, MLR outperformed these algorithms for all performance measures (see Section 3.4).

The MLR model was implemented in the glmnet R package (Friedman et al., 2010). Due to the properties described in Chapter 2, L_2 regularisation was selected and the λ value was optimised for each realisation through bootstrap analysis (25 resamples). A MLR was chosen over LASSO because it is computationally faster and easier to implement. However, in contrast to the RR in Chapter 2, MLR implements a multinomial loss function through coordinate descent.

Multinomial logistic regression can best be shown mathematically in its binomial form (Equation 4.1) which is the equivalent of the log-maximum likelihood function and therefore, the residuals should have a normal distribution (Kuhn and Johnson, 2013). However, the multinomial function does this for each soil class to form a hyperplane. Where, $L(p)$ is the binomial likelihood function or the log-

odds of the soil class occurring, λ is the degree of penalty, *P* is the number of covariates and β_j^2 is the squared coefficients.

$$logL(p) - \lambda \sum_{j=1}^{P} \beta_j^2 \tag{4.1}$$

If p is the probability of occurrence of a soil class, the binomial likelihood function is Equation 4.2. This is also known as the logit link function.

$$L(p) = log\left(\frac{p}{1-p}\right) \tag{4.2}$$

Where, p is calculated as shown in Equation 4.3. Where, β are the coefficients, x is a vector of the covariate values, and *P* is the number of covariates. This constrains the soil class probabilities from 0 to 1.

$$p = \frac{1}{1 - exp[-\beta_o + B_i x_i + \cdots . + \beta_P x_P]} \tag{4.3}$$

Besides the algorithm chosen and sampling method, this implementation of DSMART follows the process described by Odgers et al. (2014). DSMART was run with 100 realisations and area proportional sampling with a minimum of 15 and maximum of 25 random samples per polygon. The random samples were based on weights determined by the depth class probabilities specified in the LTS legend. Area proportional sampling was conducted so polygons with a larger area, also have a larger number of random samples. However, because both the LTS-GM5 and the LTS-EX5 have 12 polygons, and the LTS polygons only has three, the LTS polygons were run with a minimum of 60 samples per polygon. This was so each polygon had a similar number of resamples for each realisation. Soil depth class probabilities were calculated from counting the number of times each pixel was classified as a particular soil depth class, for all realisations. This method is similar to the method performed by Kempen et al. (2009), except the MLR was implemented through a ridge regression and the original soil legacy data was spatial scaled through TMUs before predicting soil attributes.

4.2.10 Evaluation procedure

The location of both evaluation sample designs is shown in Figure 4.3. To evaluate model performance, the depth to any root limiting layer was measured for all 93 soil observations and these observations were used to calculate the kappa coefficient, overall accuracy, producer accuracy (PA), and user accuracy (UA) of the first most probable class raster. This is a deterministic accuracy assessment of the model performance. Additionally, the models were evaluated on the combined accuracies of the two most probable class rasters. This can be seen as a stochastic accuracy assessment of model performance, where multiple realisations of soil depth classes are evaluated

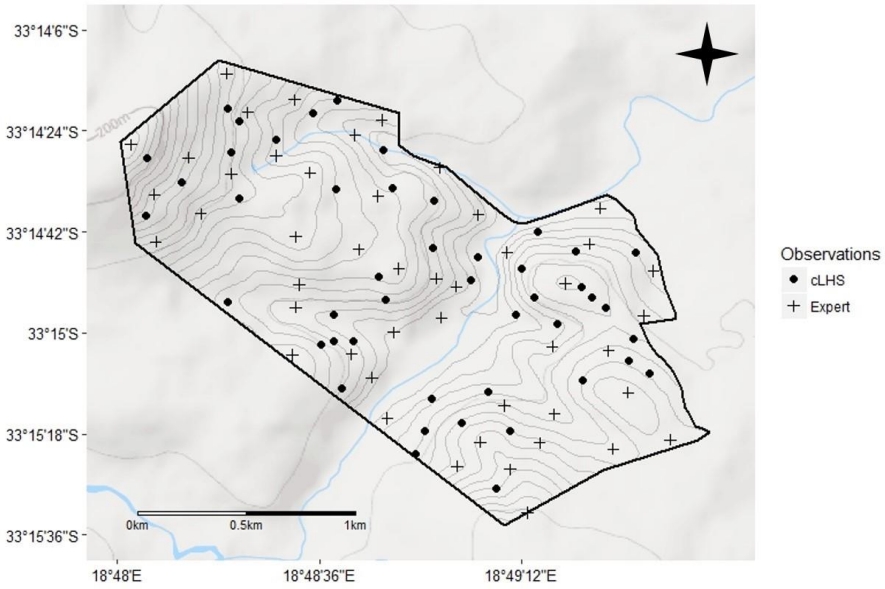


Figure 4.3: Research site with expert placed and cLHS sample locations shown within 5 m contour intervals.

To evaluate the spatial uncertainties, a confusion index (CI) between the first and second most probable soil depth class rasters was created. This follows the CI developed by Burrough *et al.* (1997). The equation for the CI is shown in Equation 4.4, where P_{max} is the probability of the most probable soil depth class and P_{max-1} is the probability of the second most probable soil depth class. This creates a CI raster where the closer a pixel is to zero, the more certain the model is of having correctly classified the soil depth class at that pixel. In other words, the larger the difference between the first and second most probable soil depth class probabilities, the more certain DSMART is of the predictions made.

4.3 Results and discussion

4.3.1 Depth class probabilities

The probabilities of each soil depth class, on each TMU, according to the LTS legend can be seen in Table 4.4. The soil depth class probabilities align with the "mental model" of the soil scientists who developed the LTS. For example, the land types show shallow soils on crest positions and deeper soils at lower elevations. Moderately deep soils do not have a clear trend for land types Ac30 and Fb97 but land type Fb96 shows moderately deep soils increasing in lower elevation positions. These probabilities are important for the final predictions as if the LTS legend is inaccurate, the final predictions will have a larger error. This was seen by Holmes et al. (2015) who stated that poorly delineated soil polygons extensively affect the outcome of the final predictions. This potential limitation can be overcome by changing the probabilities specified in the LTS legends. Vincent et al. (2018) successfully applied expert rules into the DSMART structure. However, this will require an expert pedologist familiar with a given site and places where financial resources are low, it is unlikely to find such an expert. Therefore, the probabilities were not changed as to evaluate the raw data specified in the LTS.

Table 4.4: Effective rooting depth probabilities for each terrain unit according to the LTS.

Ac30	Crest (%)	Mid-slope (%)	Foot-slope (%)	Valley (%)
Deep	0	5	40	50
Moderate	45	65	40	45
Shallow	55	30	20	5

Fb96	Crest (%)	Mid-slope (%)	Foot-slope (%)	Valley (%)
Deep	0	5	10	38
Moderate	21	32	46	47
Shallow	79	63	44	15

Fb97	Crest (%)	Mid-slope (%)	Foot-slope (%)	Valley (%)
Deep	0	7	15	95
Moderate	20	43	48	0
Shallow	80	50	37	5

4.3.2 Predicted landform elements

The proportion of TMUs predicted by the LTS-GM5 and LTS-EX5 relative to the LTS legend is shown in Figure 4.4. The LTS-GM5 underestimated crest positions by 23% while the LTS-EX5 represented this landscape position well. Mid-slope positions had the highest proportion of area for all predictions. This was expected as sloping positions are in general, the most widely seen TMU (Libohova et al., 2016; Raska, 2012). However, the LTS-GM5 overestimated mid-slope by 30% and the LTS-EX5 underestimated these positions by 47%. The LTS-GM5 underestimated foot-slope positions by 53%, and the LTS-EX5 overestimated these positions by 34%. The LTS-GM5 predicted valley positions well and the LTS-EX5 overestimated this position by 64%.

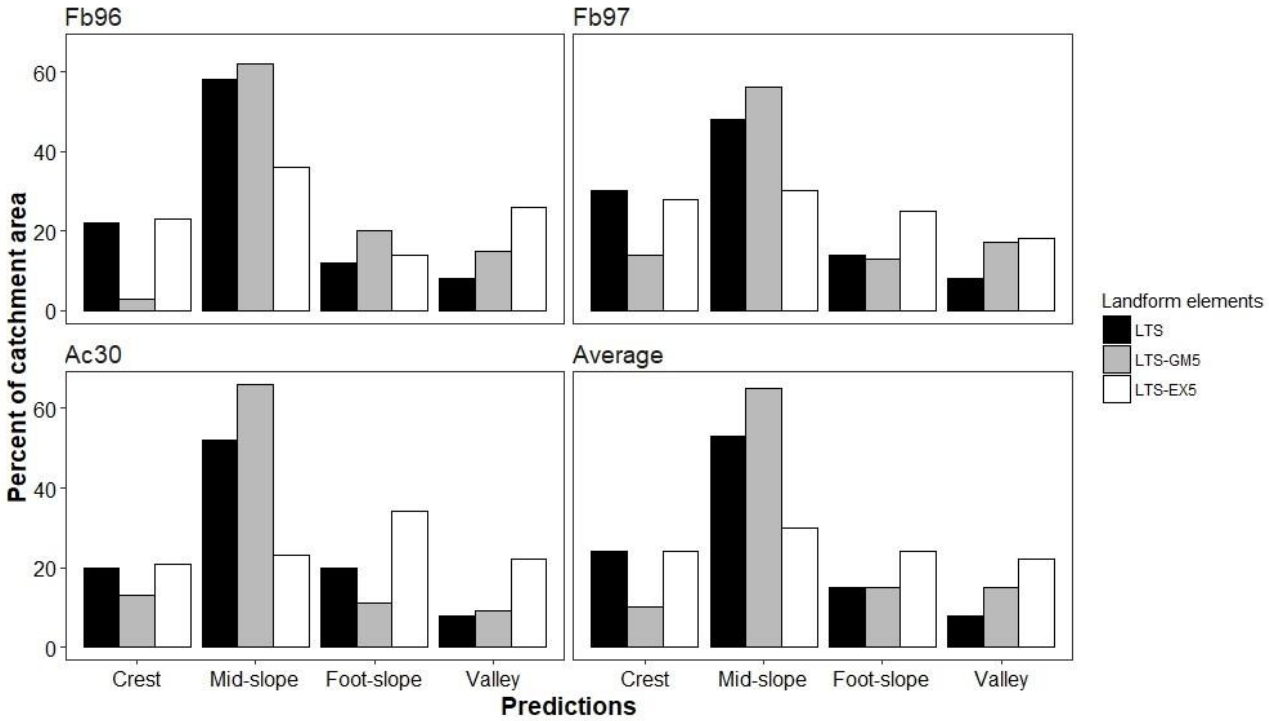


Figure 4.4: Landform element proportion of area for the LTS-GM5 and LTS-EX5 compared to the LTS.

4.3.3 Covariates and covariate importance

The covariates selected to establish soil environmental relationships and summary statistics of the covariate importance for all 100 realisations of the, LTS-GM5 (MLR) model are shown in Table 4.5. Covariate importance is defined as the scaled (0-100%) absolute value of the coefficients of the MLR model and is the average over all realisations. The covariates were selected by running DSMART and eliminating the covariates which had the lowest importance until an acceptable accuracy was achieved or the accuracy started to decrease substantially. This was a time-consuming process which

was thought to be improved by spatial principle component analysis. However, when running spatial principle component analysis to represent 95% of the covariate variability for all covariates implemented in the GSIF R package (Hengl, 2019), however, the model accuracies accuracy did not increase.

Table 4.5: The MLR model mean covariate importance for all realisations according to each depth class.

Covariates	Rank (mean)	Mean (%)	Rank (shallow)	Shallow (%)	Rank (moderate)	Moderate (%)	Rank (Deep)	Deep (%)
Convexity	1	64	1	74	3	22	1	97
Gradient	2	41	3	40	2	24	2	60
TPI	3	39	2	48	7	14	3	54
Lithology	4	27	7	19	1	36	4	28
Altitude	5	25	4	29	4	19	5	27
LS Factor	6	23	5	26	5	18	6	26
Convergence index	7	22	6	23	6	17	7	26

On average, convexity was the most important covariate followed by slope gradient, TPI, lithology, altitude, LS factor, and convergence index. Shallow and deep soils correspond to similar covariates, but for different reasons. Positions with low values for convexity and high values for topographic position and slope gradient correlate best with shallow soils. Deep soils correlate to processes controlling deposition and sedimentation, such as high convexity values and low slope gradient and TPI. Moderately deep soils correspond best to lithology; however, in general, covariates correlate less to moderately deep soils. This can be explained by moderately deep soils representing the transition from shallow to deep. This makes it more difficult for the algorithm to find covariates which separate moderately deep soils from the other soil (Burrough et al., 1997). Therefore, the covariates did not separate the classes well enough and the soil depth classes correlated to the same covariates (Chaney et al., 2016). However, when running DSMART with additional covariates, the model accuracies did not improve.

One option to improve the separation of moderately deep soils would be to run a supervised feature selection on a larger pool of covariates; however, this will increase the computation time of the algorithm and without soil point data, a supervised feature selection can be un-reliable as the random samples might not align with the feature space sufficiently. Alternatively, soil depth class criteria can be derived through a data driven technique such as k-means clustering (Forgy, 1965).

This can be seen as a pre-processing technique to find structure in the LTS legend and has been shown to improve prediction accuracy (Trivedi et al., 2015). However, this data set is too small for such clustering as Dolnicar et al. (2014) states that 70 observations are needed for sufficient cluster separation. The LTS legends for this site, only have 56 depth ranges specified.

4.3.4 Most probable class rasters

The first and second most probable class rasters produced by the LTS-GM5 through the MLR model are shown in Figure 4.5 and the probabilities of each soil depth class are shown in Figure 4.6. The first most probable class raster mimics what was described in the LTS legend. Therefore, the model extracted the expert knowledge of the soil surveyor which is important when disaggregating soil legacy data without additional knowledge of an area or in areas with limited knowledge of the spatial soil distribution. This is notable in South Africa where many soil scientists involved in creating the LTS are now retired.

Shallow soils had the highest probability on crest positions which is a result of erosional processes exceeding depositional processes (Scholms et al., 1983). Mid-slope positions had both shallow and moderately deep soils as deposition starts to increase downslope. Foot-slopes and valleys had both moderately deep soils and deep soils. Depositional processes such as colluvial wash or aeolian deposits have covered the gravelly creep and highly weathered shale, thus increasing the soil depth downslope (Lambrechts, 1983).

The first most probable class raster did not predict the shallow soils found in the middle of valleys where Typic Endoaquepts and Typic Endoaqualfs are found. Although these soils are physically deep, they have a permanent water table and therefore have a shallow depth class. An explanation for this misclassification is that shallow soils in valleys were not specified in the LTS legend. When evaluating the two most probable class rasters together, these shallow soils were still misclassified. However, these soils were classified as moderately deep soils as opposed to deep soils. This is a lesser loss of information because shallow soils are closer in taxonomical space to moderately deep soils, than deep soils are. In other words, when using soil classes, misclassification is not as consequential when the classes are similar (Rossiter et al., 2017). Minasny and McBratney (2007), show how minimising a taxonomical distance loss function in decision trees can improve accuracy of predictions. However, this has yet to be implemented in DSMART (Odgers et al., 2014).

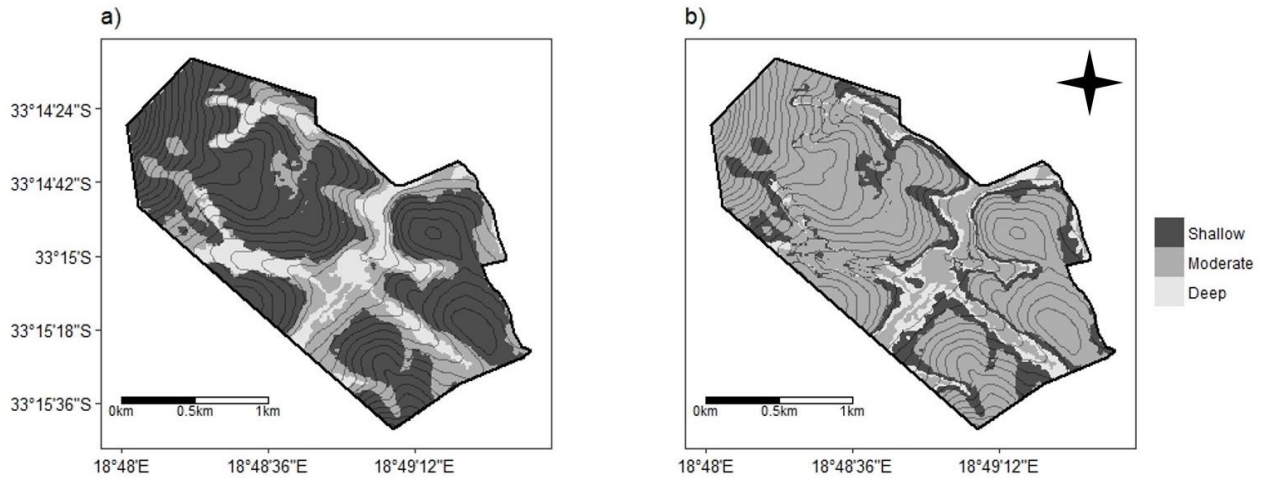


Figure 4.5: First most probable raster (a) and second most probable raster (b) of the LTS-GM5 TMUs shown with 5 m contour intervals (MLR model).

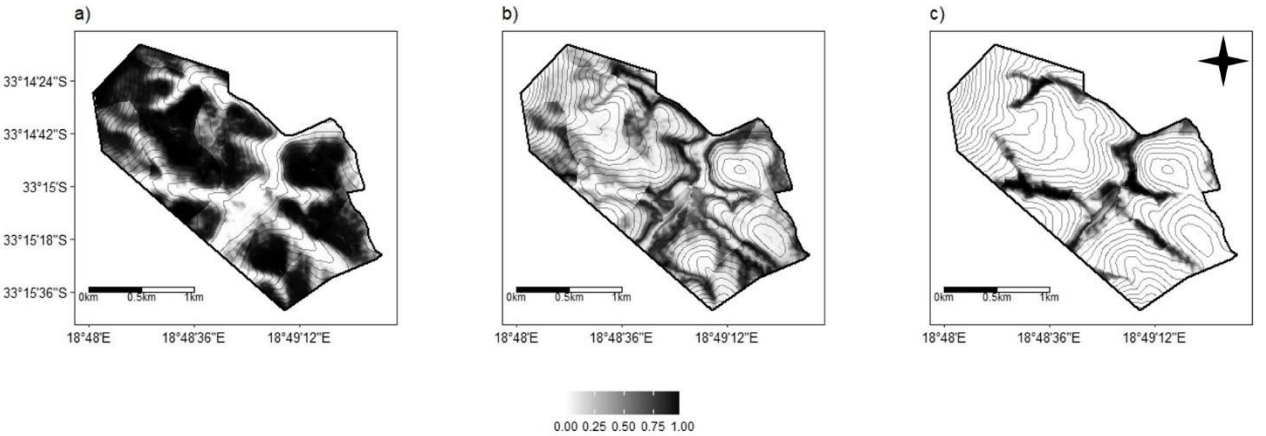


Figure 4.6: Class probability rasters for shallow (a), moderately deep (b), and deep soils (c) shown with 5 m contour intervals.

4.3.5 Accuracy assessment

The kappa coefficient and percent accuracy of the first probable class rasters, and the combined accuracy of the two most probable class rasters for all algorithms are shown in Table 4.6. What is clear from the table, is classifying TMUs (aggregated geomorphons or manual delineation) before running DSMART is required to achieve accurate results. This can be seen in the kappa coefficient of the first most probable class raster for all models run. Multinomial logistics regression is the top performing algorithm in terms of kappa coefficient and overall accuracy of the first most probable class raster for both the LTS-GM5 and LTS-EX5.

At this geographic location, spatial and soil depth class scale (3 classes), no algorithm outperform another when taking the overall accuracies from all 100 realisations according to a student T-test ($p<0.05$). Therefore, it is recommended that the simplest algorithm be optimised to increase computational efficiency. Due to the low computation demand, consistently high kappa coefficient and overall accuracy of the first most probable class rasters produced from the LTS-GM5 and LTS-EX5 models, the results produced from the MLR algorithm will be discussed further.

Table 4.6: The kappa coefficient, overall accuracy, and the combined accuracies of the first and second most probable class rasters for all algorithms achieved for the LTS-GM5, LTS-EX5, and the LTS polygons.

MLR	LTS-GM5	LTS-EX5	LTS
Kappa	0.39	0.39	-0.13
Accuracy (%)	68	68	53
Combined (%)	90	90	67
OLR	LTS_GM5	LTS-EX5	LTS
Kappa	0.27	0.31	0.00
Accuracy (%)	61	64	63
Combined (%)	80	86	67
C5.0	LTS-GM5	LTS-EX5	LTS
Kappa	0.21	0.27	0.11
Accuracy (%)	63	61	64
Combined (%)	87	92	67
RF	LTS-GM5	LTS-EX5	LTS
Kappa	0.21	0.25	0.02
Accuracy (%)	59	60	57
Combined (%)	87	89	65

The LTS-GM5 and LTS-EX5 models had similar and satisfactory results which greatly outperform the LTS model. Both models' first most probable rasters had a kappa coefficient indicating fair agreement according to Landis and Koch (1977) and accuracies similar or greater than, traditional soil map accuracies of 65% described by Marsman and de Gruiter (1986).

The satisfactory results were not expected for two reasons. First, it was assumed that the accuracies would decrease below an acceptable level (< 65%) without using soil point data to train the model,

as previously found by van Zijl et al. (2013). Secondly, the LTS was developed on a regional scale, and this study focused on the farm-scale. Therefore, the probabilities predicted from the LTS may not be represented at the farm-scale, and the variability of soil depth should increase when rescaling (McBratney, 1998b). However, this does not seem to be the case. One explanation for this could be the fact that the area is covered by not one, but three land types. It is thought that this had an averaging effect on the probabilities of soil depth classes in the LTS legend when predicting DSMART across the different land types. Furthermore, due to the soil depth classes, this effect might not be as relevant.

The increased complexity of the LTS-GM5 and LTS-EX5 also contributed to these TMUs outperforming disaggregating the LTSs polygons. A reason for this is that the polygons generated by TMUs have more detailed spatial soil information than the original land types. It is believed that this increased complexity stratifies soil depth classes into more homogenous sub-regions. This effectively aligned the random samples with the feature space which determines soil environmental relationships (Holmes et al., 2015). When the polygons are geographically large and complex (soils and terrain), such as the LTS polygons, the random samples did not align with the feature space causing less accurate predictions. This result confirms that the two-step approach where the land types are stratified with LFEs, is necessary when predicting soil depth classes at the farm-scale.

The confusion matrix and the PA and UA for the LTS-GM5 model's first most probable soil class raster is shown in Table 4.7. Shallow soils were classified with the highest accuracy followed by deep soils and moderately deep soils. These results are not surprising as shallow soils have the largest probabilities specified in the LTS legends. This was also observed in previous studies such as Holmes et al. (2015). Deep soils were predicted with high accuracy; however, these soils were overly predicted and therefore, have a low UA. Moderately deep soils were classified with the least accuracy; however, moderately deep soils have a high UA. This indicates that moderately deep soils were under predicted but the predictions made, were classified with a high accuracy. Therefore, the deterministic accuracy assessment is limited to the probabilities and scale used in the LTS. To improve the predictions of soils with a lesser probability, expert rules could be assigned for each soil depth class to the resampling procedure (Odgers et al., 2014; Vincent et al., 2018).

Table 4.7: Confusion matrix with producer accuracy (PA) and user accuracy (UA) for the first most probable class raster based on the external evaluation from 93 soil profiles.

Observed	Predicted			
	Deep	Moderate	Shallow	UA (%)
Deep	5	6	3	36
Moderate	3	10	7	50
Shallow	2	9	48	82
PA (%)	40	50	83	

When evaluating the combined accuracy of the two most probable class rasters, the LTS-GM5 model correctly classified 60% of deep, 100% of moderately deep, and 92% of shallow soils. The large increase in individual class accuracies was attributed to aggregating soil depth into only three depth classes. Therefore, the two most probable class rasters account for almost all of the soil class variability. If the number of soil depth classes were to increase, individual soil class accuracy would not increase as much. Additionally, these results should be interpreted with caution, especially where there is a high CI (e.g., moderately deep soils).

Shallow and deep soil class accuracy increased by around 20% when assessing the accuracy of the combined rasters; however, moderately deep soil class accuracy increased by 50%. This can be attributed to the model overly predicting shallow and deep soils in the first most probable class raster and underestimating these soils in the second most probable class raster. Moderately deep soils showed the opposite trend leading to the large increase in moderately deep soils accuracy when combining the two most probable class rasters. Therefore, DSMART's ability to produce multiple probability class rasters may improve results when downscaling to the farm-scale relative to deterministic approaches.

4.3.6 Spatial uncertainties

The CI for the LTS-GM5 is shown in Figure 4.7. The CI is notable because it gives the uncertainties of soils in a spatial context which could assist in additional soil surveys and give insight into model performance (Odgers et al., 2014). The higher the CI, the more uncertain the model is of the soil depth class predicted at that given pixel. The LTS-GM5 had an average CI of 0.25 which is a slight improvement from the LTS-EX5 of 0.27 and a large improvement compared to the LTU of 0.57. The

low average uncertainty suggests that the LTS-GM5 MLR model was the appropriate algorithm for this site. When running DSMART with OLR, C5.0 algorithm, and RF, the average CI increased to 0.32, 0.56, and 0.35 for the LTS-GM5, respectively.

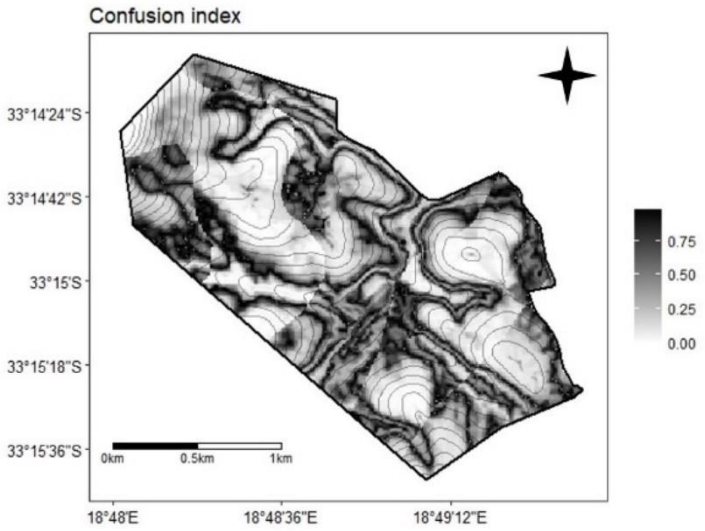


Figure 4.7: Confusion index between the first and second most probable class rasters shown with 5 m contours.

It is evident that the uncertainties are spatially autocorrelated and that there are patterns associated with soil depth class. Shallow soils had an average uncertainty of 0.15 and had the lowest uncertainty of any soil depth class. Moderately deep soils had the highest uncertainty of any soil depth class with an average of 0.64. Deep soils had an average uncertainty of 0.48. It was thought that the more accurate the model classified a soil depth class, the lower the uncertainties would be. This was the case when evaluating the first most probable raster, however, this was not the case when evaluating the two most probable rasters. For example, moderately deep soils had the highest uncertainty, however, when evaluating the two most probable rasters, moderately deep soils were correctly classified 100% of the time. An explanation for this is the weaker trend in the LTS data for moderately deep soils leading to the high uncertainties, although they were classified with high accuracy. Therefore, the spatial distribution of moderately deep soils is associated with a high uncertainty.

4.4 Conclusion

This study demonstrates a two-step disaggregation approach of a national resource inventory into a soil depth class map at the farm-scale. Landform elements were classified through the geomorphon algorithm and overlaid with the LTS. This created polygons with unique distribution of soil attributes.

The polygons were used for resampling in DSMART implemented through a MLR to predict soil depth classes. This approach was compared with a manually delineated TMUs as resample polygons and disaggregating the LTS without TMU classification. The main findings of this study are:

- A soil depth class map was produced from a national resource map and achieved an accuracy similar or greater than, conventional soil survey accuracies.
- To disaggregate the LTS into farm-scale soil depth classes, initial TMU classification is required to produce polygons from which, DSMART can resample.
- Aggregated geomorphons can be used to predict TMUs specified in the LTS and therefore, can produce the first step and increase efficiency of the disaggregation process.
- Multinomial logistics regression implemented through DSMART, is capable of extracting soil depth classes from the probabilities specified in the LTS legend.
- The trend in the LTS legend greatly affects model performance and may be improved through incorporating expert knowledge of both TMUs and soils of the area.

This study has shown the potential to increase the accessibility and interpretability of the LTS. However, there is much work that needs to be done in terms of geomorphons and DSMART. This approach needs to be implemented in different geographic regions, on different scales, and with additional soil attributes. This may need to be addressed through standardising the implementation of this approach. A feature selection algorithm to select both geomorphons and covariates from the probabilities in the LTS may be an option. Then it has the potential to disaggregate the LTS across South Africa into soil maps that can be used for a variety of purposes.

4.5 References

Breiman, L., 2001. Random Forests. Berkeley, California. https://doi.org/10.1017/CBO9781107415324.004

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

Bunning, S., McDonagh, J., Rioux, J., 2011. Part 2 - Field methodology and tools, in: Manual for Local Level Assessment of Land Degradation, Sustainable Land Management and Livelihoods. FAO, Roam, p. 60.

Burrough, P.A., Van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: Spatial correlation, confusion and boundaries. Geoderma 77, 115–135. https://doi.org/10.1016/S0016-7061(97)00018-9

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analysis (SAGA). Geoscientific Model Development.

https://doi.org/doi:10.5194/gmd-8-1991-2015

Devia, G.K., Ganasri, B.P., Dwarakish, G.S., 2015. A Review on Hydrological Models. Aquat. Procedia 4, 1001–1007. https://doi.org/10.1016/j.aqpro.2015.02.126

Dolnicar, S., Grün, B., Leisch, F., Schmidt, K., 2014. Required Sample Sizes for Data-Driven Market Segmentation Analyses in Tourism. J. Travel Res. 53, 296–306. https://doi.org/10.1177/0047287513496475

Dragut, L., 2011. Automated classification of topography from SRTM data using object-based image analysis. Geomorphometry 141–12.

Forgy, E., 1965. Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification. Biometrics 21, 768–769.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1–24. https://doi.org/10.18637/jss.v033.i01

GRASS Development Team, 2017. Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Found.

Grunwald, S., 2006. Environmental Soil-Landscape Modeling. Taylor & Francis, Boca Raton, Florida.

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma 185–186, 37–47. https://doi.org/10.1016/j.geoderma.2012.04.001

Hengl, T., 2019. GSIF: Global Soil Information Facilities. R package version 0.5-5.

Hengl, T., Evans, I.S., 2009. Mathematical and digital models of the land surface. Dev. Soil Sci. 33, 31–63. https://doi.org/10.1016/S0166-2481(08)00002-0

Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. CSIRO 53, 865–880.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25, 295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Irvin, B.J., Ventura, S.J., Slater, B.K., 1997. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. Geoderma 77, 137–154. https://doi.org/10.1016/S0016-7061(97)00019-0

Iwahashi, J., Pike, R.J., 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology 86, 409–440. https://doi.org/10.1016/j.geomorph.2006.09.012

Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons - a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156. https://doi.org/10.1016/j.geomorph.2012.11.005

Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands. Geoderma 241–242, 313–329. https://doi.org/10.1016/j.geoderma.2014.11.030

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach, Geoderma. Elsevier B.V. https://doi.org/10.1016/j.geoderma.2009.04.023

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, the R.C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., 2018. Caret: Classification and

Regression Training.

Lambrechts, J.J.N., 1983. Soils, Soil Process and Distribution in the Fynbos Region: An Introduction. Council for Scientific and Industral Research, Pretoria.

Land Type Survey Staff, 1972-2006. Land Types of South Africa on 1:250 000 scale. Pretoria, South Africa.

Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33, 159–174.

Libohova, Z., Winzeler, H.E., Lee, B., Schoeneberger, P.J., Datta, J., Owens, P.R., 2016. Geomorphons: Landform and property predictions in a glacial moraine in Indiana landscapes. Catena 142, 66–76. https://doi.org/10.1016/j.catena.2016.01.002

MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets Syst. 113, 81–109. https://doi.org/10.1016/S0165-0114(99)00014-7

Malone, B.P., Styc, Q., Minasny, B., McBratney, A.B., 2017. Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. Geoderma 290, 91–99. https://doi.org/10.1016/j.geoderma.2016.12.008

Marsman, B.A., De Gruijter, J.J., 1986. Quality of Soil Maps, in: Soil Survey Papers. Netherlands Soil Survey Institute, Wageningen, The Netherlands.

McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. Nutr. Cycl. Agroecosystems 50, 51–62. https://doi.org/10.1023/A:1009778500412

McBratney, A.B., Santos, M.L.M., Minasnya, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142, 285–293. https://doi.org/10.1016/j.geoderma.2007.08.022

Myburgh, P.A., van Zijl, J.L., Condradie, W.J., 1996. Effect of Soil Depth on Growth and Water Consumption of Young Vitis vinifera L . cv . Pinot noir. South African J. Enol. Vitic. 17, 53–62.

Nauman, T.W., Thompson, J.A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385–399. https://doi.org/10.1016/j.geoderma.2013.08.024

Nauman, T.W., Thompson, J.A., Rasmussen, C., 2014. Semi-Automated Disaggregation of a Conventional Soil Map Using Knowledge Driven Data Mining and Random Forests in the Sonoran Desert, USA. Photogramm. Eng. Remote Sens. 80, 353–366. https://doi.org/10.14358/PERS.80.4.353

Odgers, N., Malone, B.P., 2017. rdsmart: Disaggregation and harmonisation of soil map units through resampled classification trees (R package version 2.0.3).

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100. https://doi.org/10.1016/j.geoderma.2013.09.024

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc, San Francisco, California.

R Core Team, 2017. R: A language and environment for statistical computing.

Raska, P., 2012. Biogeomorphologic Approaches to a Study of Hillslope Processes Using Non-Destructive

Methods, in: Studies on Environmental and Applied Geomorphology. Elsevier Science B.V., Amesterdam, pp. 21–41.

Ray, S.S., Singh, J.P., Das, G., Panigrahy, S., 2004. Use of high resolution remote sensing data for generating site-specific soil mangement plan, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Rondeaux, G., Steven, M., Baret, F., 1996. Optimization of soil-adjusted vegetation indices. Remote Sens. Environ. 55, 95–107. https://doi.org/10.1016/0034-4257(95)00186-7

Rossiter, D.G., Zeng, R., Zhang, G.L., 2017. Accounting for taxonomic distance in accuracy assessment of soil class predictions. Geoderma 292, 118–127. https://doi.org/10.1016/j.geoderma.2017.01.012

Rozanov, A., Lessovaia, S., Louw, G., Polekhovsky, Y., de Clercq, W., 2017. Soil clay mineralogy as a key to understanding planation and formation of fluvial terraces in the South African Lowveld. Catena 156, 375–382. https://doi.org/10.1016/j.catena.2017.04.027

Sarmento, E.C., Giasson, E., Webster, E.J., Flores, C.A., Hasenack, H., 2017. Regional Disaggregating conventional soil maps with limited descriptive data: A knowledge-based approach in Serra Gaúcha , Brazil. Geoderma Reg. 8, 12–23. https://doi.org/10.1016/j.geodrs.2016.12.004

Schmidt, J., Hewitt, A., 2004. Fuzzy land element classification from DTMs based on geometry and terrain position. Geoderma 121, 243–256. https://doi.org/10.1016/j.geoderma.2003.10.008

Scholms, B.H.A., Ellis, F., Lambrechts, J.J.N., 1983. Soils of the Cape Coastal Platform, in: Deacon, H.J., Hendey, Q.B., Lambrechts, J.J.N. (Eds.), . Council for Scientific and Industral Research, Pretoria.

Scull, P., Franklin, J., Chadwick, O. a., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27, 171–197. https://doi.org/10.1191/0309133303pp366ra

Silva, S.H.G., Menezes, M.D., Mello, C.R., Góes, H.T.P., Owens, P.R., Curi, N., 2016. Geomorphometric tool associated with soil types and properties spatial variability at watersheds under tropical conditions. Sci. Agric. 73, 363–370. https://doi.org/10.1590/0103-9016-2015-0293

Silva, S.H.G., Owens, P.R., Duarte de Menezes, M., Reis Santos, W.J., Curi, N., 2014. A Technique for Low Cost Soil Mapping and Validation Using Expert Knowledge on a Watershed in Minas Gerais, Brazil. Soil Sci. Soc. Am. J. 78, 1310. https://doi.org/10.2136/sssaj2013.09.0382

Smith, S., Bulmer, C., Flager, E., Frank, G., Filatow, D., 2010. Digital soil mapping at multiple scales in British Columbia, Canada. Progr. Abstr. 4th Glob. Work. Digit. Soil Mapp. 17.

Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334–345. https://doi.org/10.1016/j.geoderma.2013.08.018

Trivedi, S., Pardos, Z.A., Heffernan, N.T., 2015. The Utility of Clustering in Prediction Tasks, arXiv.

Van Zijl, G.M., Le Roux, P.A., Turner, D.P., 2013. Disaggregation of land types using terrain analysis, expert knowledge and GIS methods. South African J. Plant Soil 30, 123–129. https://doi.org/10.1080/02571862.2013.806679

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2018. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142. https://doi.org/10.1016/j.geoderma.2016.06.006

Wiese, L., Ros, I., Rozanov, A., Boshoff, A., de Clercq, W., Seifert, T., 2016. An approach to soil carbon accounting and mapping using vertical distribution functions for known soil types. Geoderma 263, 264–273. https://doi.org/10.1016/j.geoderma.2015.07.012

Yang, L., Jiao, Y., Fahmy, S., Zhu, A.X., Hann, S., Burt, J.E., 2011. Updating Conventional Soil Maps through Digital

Soil Mapping. Soil Sci. Soc. Am. 75, 1044–1053. https://doi.org/10.2136/sssaj2010.0002

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

Chapter 5 Comparing algorithms to disaggregate complex soil polygons in contrasting environments

This chapter is based on the publication Flynn, T., van Zijl, G., van Tol, J., Botha, C., Rozanov, A., Warr, B., Clarke, C. 2019. Comparing algorithms to disaggregate complex soil polygons in contrasting environments. *Geoderma* 352, 171 – 180 found in Appendix C.

Abstract:

In South Africa, the only soil resource available with full spatial coverage is the LTS. Disaggregating this polygon-based inventory, is thus a logical step to create more detailed soil maps covering the entire country. The polygons are large in area encompassing complex soil-terrain patterns and research into disaggregation techniques has been limited. This study aimed to compare 10 algorithms, implemented through a modified DSMART model, in their ability to disaggregate two polygons into soil associations in two environmentally contrasting locations. One site had high relief and strong catenal sequences (eastern KwaZulu-Natal Province) and the other site had low relief and a strong geological control of soil types (northern Eastern Cape Province). The algorithms compared were based on previous studies which included *k*-nearest neighbour, nearest shrunken centroid, discriminant analysis, multinomial logistics regression, linear and radial support vector machines, decision trees, stochastic gradient boosting, random forest, and neural networks. The method involves stratifying the polygons with landform elements, randomly sampling the landform elements, allocating the soil classes based on the resource inventory, and predicting soil associations across a stack of covariates. This was done in an iterative process, creating multiple realisations of the soil distribution. The performance of each algorithm was based on their kappa and uncertainties. It was found that in general, robust linear models which either utilise an embedded feature selection or regularise covariates, performed best. In the area with high relief and clear toposequences, nearest shrunken centroid was the top performing algorithm with a kappa of 0.42 and an average uncertainty of 0.22. In the area with relatively low relief and complex geology, the results were unsatisfactory. However, a regularised multinomial regression was the top performing algorithm, achieving a kappa of 0.17 and an average uncertainty of 0.84. The results of this study highlight the versatility of a technique to disaggregate South Africa's national resource inventory, where algorithms can be chosen on expert knowledge, model averaging can be performed, the top performing algorithm can be chosen, and algorithm parameters can be optimised.

5.1 Introduction

There have been studies that compare several algorithms trained on point observations to predict soil type (Brungard et al., 2015; Heung et al., 2016). In contrast, there has been little to no research comparing algorithms used to disaggregate national resource inventories in different environmentally contrasting areas. This is surprising as these inventories are seen as a wealth of information that often cover much, if not all, of a country.

Decision tree based approaches have become popular in disaggregation models (Bui and Moran, 2001; Nauman and Thompson, 2014; Silva et al., 2016; Subburayalu et al., 2014). This is because decision tree algorithms can imitate the "mental model" of soil scientists and can handle either discrete or continuous data (Bui et al., 1999; Bui and Moran, 2001; Odgers et al., 2014). Decision trees can also handle non-linear relationships making them powerful classifiers (Breiman et al., 1984). One criticism of decision trees is that they are subject to overfitting (Grunwald, 2009) and therefore, other studies such as Häring et al. (2012), Nauman et al. (2014), Chaney et al. (2016), Vincent et al. (2016), and Møller et al. (2019) have used Random Forest (RF; Breiman, 2001). Random Forest is also known for increasing model performance by reducing the variance of predictions through its ensemble approach (Bühlmann and Yu, 2002; Strobl et al., 2009). However, it was found that, when implemented through DSMART, a regularised multinomial logistics regression performed better than decision trees and RF when disaggregating LTS polygons at the farm-scale into a depth class map in South Africa as shown in Chapter 4.

Soil information in the LTS is strongly tied to the influence each TMU exerts on soil formation. At some spatial scales and locations, this is logical as TMUs distinguish the boundaries between processes such as accumulation, deposition, and leaching potential (Evans, 2012b). Therefore, both conceptual and DSM techniques have focused on utilising these relationships when disaggregating the LTS. However, this approach is problematic when topography is not the main driver of soil formation. For example, when land types cross many contrasting parent materials, parent material will exert a strong influence on soil formation as it affects both physical and chemical soil properties (Jenny, 1941).

Studies into disaggregating the LTS include van Zijl et al. (2013), who disaggregated two land types in KwaZulu-Natal Province through an expert rules system and SoLIM software (Zhu, 1997). The

authors concluded that a field survey was required to disaggregate the LTS as the disaggregation technique achieved a 35% accuracy. When adding observations to the model, the results drastically improved. The authors also concluded that adding lithology to the rules increased map usability but not map accuracy. Botha (2016), also used an expert knowledge approach whereby, TMUs were classified and assigned a dominant soil type based on the LTS data. The results were satisfactory for a high relief area (88%) but were un-satisfactory for an area controlled more by geology (37%). In Chapter 4, the LTS was disaggregated into soil depth classes at the farm-scale by stratifying LTS polygons with TMUs as an input map for DSMART. The study achieved a satisfactory accuracy of 68 to 90% depending on how many probability class rasters were used for evaluation.

Other studies that disaggregate complex soil-terrain polygons include Bui and Moran (2001), who used *k*-means clustering to classify soil associations and decision trees to classify fluvial facies which were strongly correlated to soil texture. The authors achieved an accuracy of 76 to 83% depending on the site in western New South Wales, Australia. Holmes et al. (2015), disaggregated soil-terrain polygons through DSMART in the whole of Western Australia. The authors implemented C4.5 decision trees and achieved an accuracy of 40% according to the three most probable class rasters and achieved a 71% accuracy when using higher levels of the soil classification system. However, these studies were conducted on much larger areas than the conventional approach in South Africa conventionally focused on disaggregating a single land type for a specific region.

The objective (Framework 2, Objective 4) of this study was to evaluate 10 algorithms on their ability to disaggregate the LTS in two environmentally contrasting areas (high relief, strong catenal sequence vs low relief, weak catenal sequence, and strong soil-geological relationships) using a modified DSMART model. The model allows the implementation of many classifiers and incorporates additional features that can be used to optimise the model for an area. This also has implications for disaggregating large datasets such as SOTER (Soil Terrain Dataset) (Dijshoon et al., 2008). Therefore, it has implications for further work in larger areas. This can be seen as an add-on to DSMART which can also be applied over different scales.

5.2 Method and materials

5.2.1 Site description

Two land types were used in this study (Figure 5.1), one site at Cathedral Peak in eastern KwaZulu-Natal Province (28˚ 30' S to 29˚ 30' S and 29˚ 00' E to 29˚ 30' E) and another site at Ntabelanga in northern Eastern Cape Province (31˚ 03' S to 29˚ 09' S and 28˚ 30' E to 28˚ 44' S). Both sites were selected due to their contrasting environments and data availability.

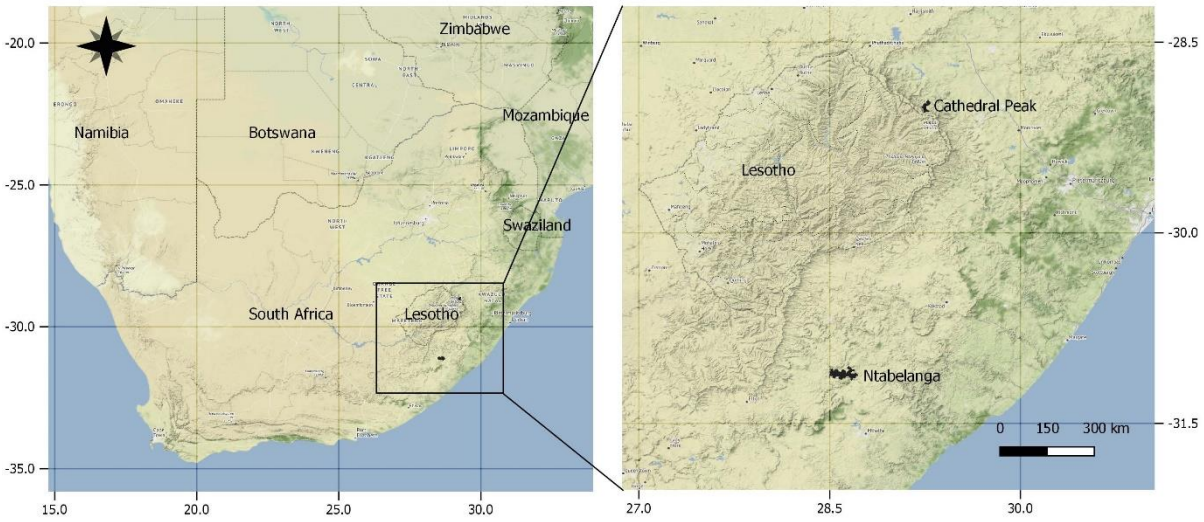


Figure 5.1: The two study sites within southern Africa and zoomed into the eastern region of South Africa.

The Cathedral Peak site forms part of the South African National Environmental Network (SAEON) and consists of several protected, near pristine catchments. The site is located in the Drakensberg mountain range close to the border of Lesotho. It is the Ac265 land type (sheet 2828 Harrismith) encompassing 9.5 km^2 of relatively high relief and uniform geology comprising of basaltic rocks of the Drakensberg formation (Land Type Survey Staff, 1972-2006). Cathedral Peak has an Ustic climate with an average annual precipitation of 1,130 mm (Schulze, 2007). The altitude of the study area, ranges from 1,827 to 2,068 m and is mainly covered by mesic grasslands interspersed with forest patches and wetlands.

The Ntabelanga site is the Db334 land type (sheet 3128 Umtata) encompassing 7.4 km^2 of relatively low relief ranging in altitude from 871 m to 1128 m with a complex geology. The geology consists of brownish-red and grey mudstone and sandstone of the Tarkastad Subgroup, Beaufort Group with dolerite intrusions (Land Type Survey Staff, 1972-2006). This area was part of the old Transkei Homeland. Agricultural production is classified as Maize Mixed Farming (Dixon et al., 2001) on state-

owned land administered through the Tribal Authority system. Soils in the area are extremely susceptible to erosion, yet the site is earmarked for construction of the large multipurpose storage Ntabelanga dam in the Tsitsa River (Van Tol et al., 2014). The Ntabelanga area has a semi-arid climate with an average annual precipitation of 700 mm.

5.2.2 Land type terrain data

A 2-dimensional depiction of the manually delineated TMUs on each land type are shown in Figure 5.2. Cathedral Peak is dominated by mid-slope (3) consisting of 85% of the area, crest (1), and valley positions (5) make up 10% and 5% of the area, respectively. The slope at Cathedral Peak ranges from 4 to 60%. Ntabelanga is dominated by foot-slopes (4), encompassing 80% of the area, while mid-slope and valley positions consist of 10% of the area in total. The slope on the land type ranges from 4 to 10%. (Land Type Survey Staff, 1972-2006).

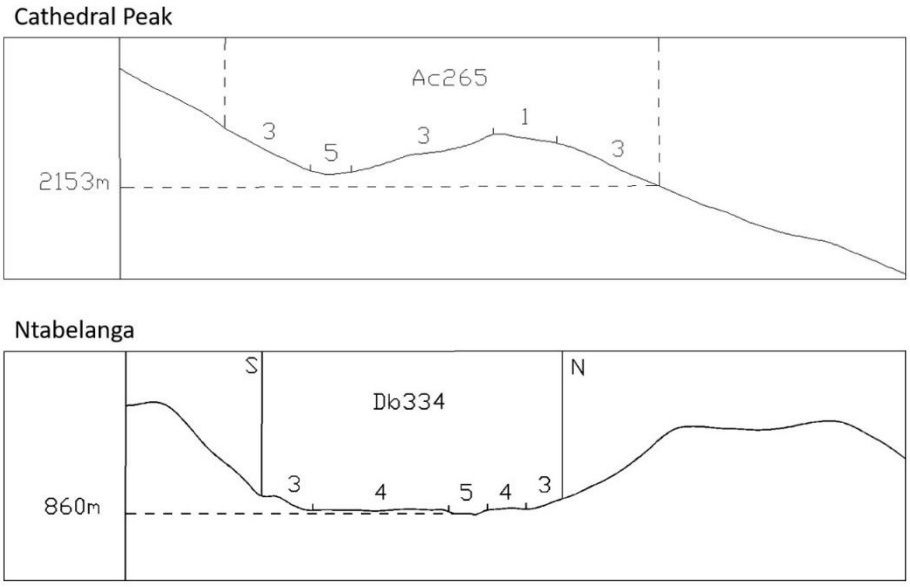


Figure 5.2: Terrain morphological units (TMUs) situated on Cathedral Peak (Ac265) and Ntabelanga (Db334) taken from (Land Type Survey Staff, 1976-2002).

5.2.3 Land type soil data

Cathedral Peak has a clear toposequence and therefore, soil associations were classified according to the toposequences. The soils were aggregated into three associations consisting of a shallow association (Lithic Haplustepts and Lithic Humustepts), characterised by soils grading into bedrock within 50 cm from the soil surface on crest positions, an apedal association (Lithic and Typic Haplustox), comprised of generally deep soil with an apedal structure on mid-slopes, and a wet

association (Typic Haplohumist), soils showing morphological signs of gleying in the valleys. Therefore, the soils were aggregated based on lithic, oxic, and hydromorphic soil properties.

Ntabelanga's soils are more structured, complex, and are heavily controlled by geology/lithology. The soils were aggregated into apedal (Typic/Plinthic Haplustox), duplex (Typic Albaqualfs), semi-duplex or pedo (Typic Haplustalfs), shallow (Lithic Haplustept and Humustepts) and wet (Endo Aqualfs and Aquepts) associations. Soil associations were based on erosion potential. For example, apedal soils have a deep profile and high iron content making them resistant to erosion. In contrast, duplex soils have a discrete textural boundary due to their binary profile making them highly prone to crusting and erosion. The duplex and pedo soil associations differ in that the duplex soil association comprises soils with a prismatic structure in the subsoil, while the subsoil structure is angular blocky in the pedo soil association. As the names suggest, the shallow soil association grades into bedrock within 50 cm of the soil surface, while wet soil association shows morphological evidence of water logging. The percent area of each soil association on the TMUs are shown in Table 5.1.

Table 5.1: Soil associations percent area on each TMU for Cathedral Peak and Ntabelanga.

Cathedral Peak	Associations	Crest (%)	Mid-slope (%)	Valley (%)
	Apedal	-	60	-
	Shallow	100	40	-
	Wet	-	-	100
Ntabelanga	Association	Mid-slope (%)	Foot-slope (%)	Valley (%)
	Apedal	10	15	-
	Duplex	-	15	20
	Pedo	30	40	40
	Shallow	60	10	5
	Wet	-	15	35

5.2.4 Polygon stratification

A topographic index landform classification (TPIc) was used to predict LFEs on each land type. The TPIc system compares the elevation at each pixel to the neighbourhood around that pixel (Weiss, 2000). For this study, the algorithm was implemented with a neighbourhood of 100 m for both land types. The TPIc and neighbourhood size were used to compare DSMART with an expert driven approach by Botha (2016) at the same locations and who used the same TPIc.

The TPIc units were aggregated into TMUs according to Table 5.2. The TMU development is in contrast to Chapters 3 and 4, where geomorphons were used to stratify the landscape. Additionally, in contrast to the previous work, this study focused on a regional scale. The aggregation was a subjective procedure based on the LTS specifications. For example, there is no foot-slope TMU at Cathedral Peak and no crest TMU classified on the Ntabelanga land type. Therefore, each land type was stratified into three polygons as two TMUs are not present at both sites.

Table 5.2: Aggregation of TPIc landform elements into TMUs for Cathedral Peak and Ntabelanga.

TPIc	Cathedral Peak	Ntabelanga
9	Crest	Mid-slope
8	Crest	Mid-slope
7	Crest	Mid-slope
6	Mid-slope	Foot-slope
5	Mid-slope	Foot-slope
4	Mid-slope	Foot-slope
3	Mid-slope	Foot-slope
2	Mid-slope	Foot-slope
1	Valley	Foot-slope
0	Valley	Valley
-1	Valley	Valley

5.2.5 Model training

The algorithms were trained and predicted using a modified DSMART model in R software (R Core Team, 2017). The modified method incorporates the caret R package where different classification algorithms can be used (Kuhn et al., 2018). The caret R package also allows for optimisation such as cross-validation, different sampling techniques such as up-sampling, and pre-processing such as centring and scaling of the covariates. The R software code developed, can be found on GitHub. Besides the incorporation of the caret R package, the model is similar to that of the rdsmart package (Odgers and Malone, 2017). For *n* realisations, the modified DSMART model is as follows:

1. Stratify the LTS with TMUs and prepare covariates.
2. Draw *m* random samples from each TMU.
3. Assign samples a soil type based on probabilities specified in the LTS.
4. Train an algorithm on covariates (caret R package).
5. Predict soil type across covariates.

Once *n* realisations have been trained and predicted:

6. Count number of times each pixel is classified a soil type.
7. Calculate probabilities based on counts (counts/total).
8. Determine soil type at each pixel.

In this study, the method follows that of Chapter 4, 15 random samples (*m*) were drawn for each TMU, assigned a soil association according to the LTS legend, covariate values were extracted, and soil associations were predicted for 100 realisations (*n*). However, the resampling procedure differs from that of Chapter 4, as the soil associations with the least probability (in the TMUs) were up-sampled after drawing the 15 random samples. Up-sampling randomly samples with replacement the soil classes with low probabilities to get as many samples as the dominant soil class. This a type of bootstrapping to account for soil class imbalances on each TMU.

The soil class assignment during the resampling procedure can be seen as a target-based approach on landscape rules (Odgers et al., 2014). A target-based approach tries to resample the polygons and assign soil classes based on known soil-environmental relationships. This target-based approach is in contrast to other methods such as Häring et al. (2012) and Vincent et al. (2016), who used landscape rules to assign samples a soil type. Additionally, this approach is different than that of Møller et al. (2019), as it only uses landscape rules found in the original resource inventory.

Model development was an iterative process using 10 different algorithms shown in Table 5.3. The algorithms were largely selected based on the studies by Brungard et al. (2015) and Heung et al. (2016), who compared similar algorithms in three semi-arid regions in western USA and British Columbia, Canada, respectively. For detail into each algorithm, see Hastie et al., (2009) and Kuhn and Johnson, (2013).

Table 5.3: Classification algorithms used to predict soil associations at Cathedral Peak and Ntabelanga.

Algorithm	Type
k-nearest neighbour (KNN)	Distance based learner
Nearest shrunken centroid (NSC)	Distance based learner
Linear discriminatory analysis (LDA)	Simple linear model
Multinomial ridge regression (MLR)	Generalised linear model (L_2regularised)
C5.0 decision trees (C5)	Tree based learner
Random forest (RF)	Multiple decision trees grown in parallel
Stochastic gradient boosting (SGB)	Multiple decision trees grown in sequence
Linear support vector machines (SVL)	Linear boundary learner
Radial support vector machines (SVR)	Radial boundary learner
Multilayer perceptron (MLP)	Multiple hidden layer neural network

5.2.6 Covariates

Topographic covariates at each site were developed from a 30 m Advanced Land Observation digital elevation model (DEM). The resolution of the DEM was used to define the predictions final resolution. The covariates used to train the models were altitude, aspect, catchment area and slope, convexity, downslope (DC) and upslope curvature (UC), plan curvature, profile curvature, local curvature (LC), LS factor, multiresolution valley bottom flatness (MRVBF), negative openness (NO), SAGA wetness index (SWI), sky view factor, slope, terrain factor, and terrain roughness. All topographic covariates were developed in the System for Automated Geoscientific Analysis (Conrad et al., 2015). These covariates were thought to describe the topography of both land types sufficiently.

In addition to topographic covariates, spectral covariates were developed at Ntabelanga from the Sentinel 2A satellite (11th November 2018) and were mean aggregated into a 30 m resolution. The addition of spectral covariates was done with the knowledge that the soils are controlled less by topography. The spectral bands and indices can be seen in Table 5.4. These covariates were thought to represent soil, vegetation, and parent material according to the *scorpan* method (McBratney et al., 2003). A description of the spectral indices can be found in Bannari et al. (1995) and Ray et al. (2004).

Table 5.4: Spectral covariates obtained and developed at Ntabelanga.

Bands	Band origin (μm)	Symbol
Blue	0.490	B
Green	0.560	G
Red	0.665	R
Near infrared (NIR)	0.842	NIR

Indices	Equation	Property
Brightness Index (BI)	$\frac{R^2 + G^2 + B^2}{3^{0.5}}$	Reflectance
Coloration Index (CI)	$(R - G)/(R + G)$	Soil colour
Redness Index (RI)	$R\hat{}2/(B * G\hat{}3)$	Hematite
Saturation Index (SI)	$(R - B)/(R + B)$	Spectral slope
NDVI	$(NIR - R)/(NIR + R)$	Chlorophyll

5.2.7 Field observations

Field observations at Cathedral Peak and Ntabelanga were conducted during previous studies detailed below and shown in Figure 5.3 and Figure 5.4, respectively. Soils in both surveys, were classified according to South African Soil Taxonomy (Soil Classification Working Group, 1991). It should be noted that the original LTS soil profiles were not considered in either land type as they were not specified in the LTS legend.

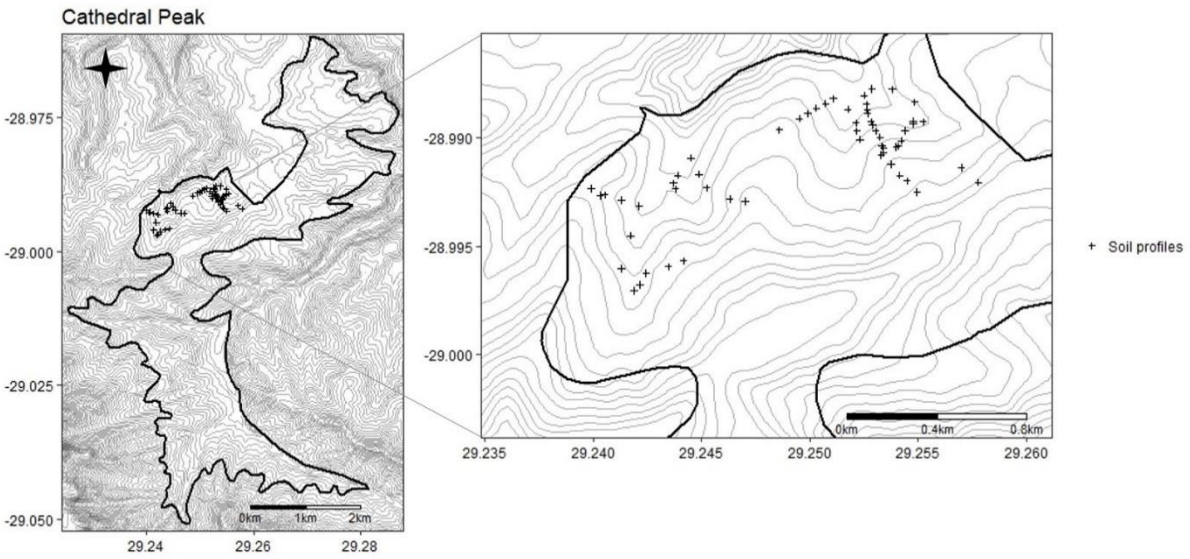


Figure 5.3: Fifty-eight soil profiles in the Cathedral Peak land type (Ac265) shown on 20 m contour intervals.

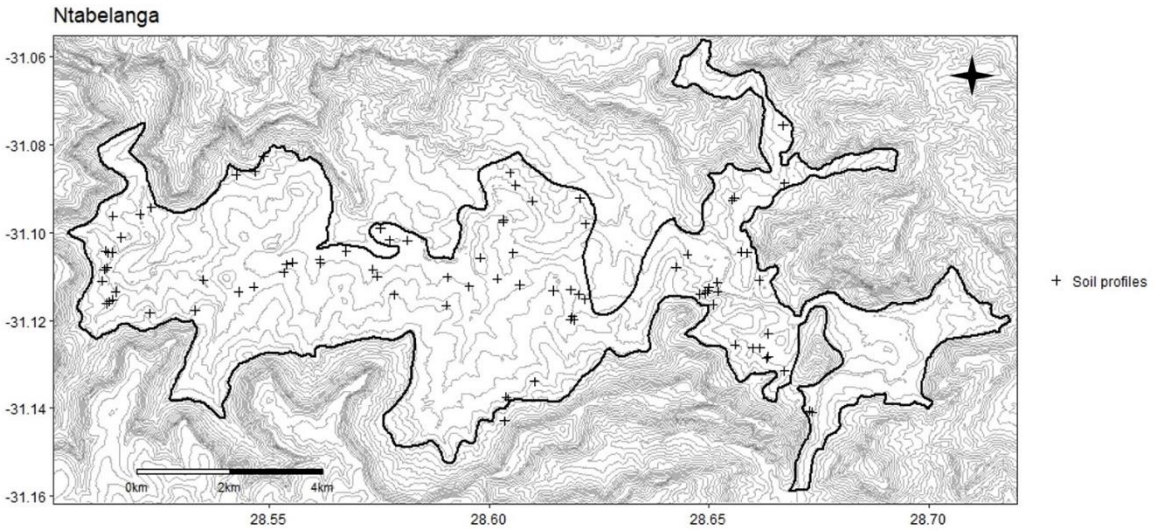


Figure 5.4: Eighty-seven soil profiles in Ntabelanga land type (Db334) shown on 20 m contour intervals.

At Cathedral Peak, 58 stratified random samples were targeted which were stratified between the Cathedral Peak research catchments within the land type (van Zijl and Botha, 2016). The profiles were classified by 50 expert participants from the South African Soil Surveyor's Organization (SASSO) working group. Even though the samples were clustered within the research catchments, they are deemed to sufficiently represent the land type for evaluation of the models, as the soil distribution should be similar throughout the land type.

Eighty-seven soil profiles were classified and sampled in the Ntabelanga area as part of three projects. The first was to characterise the erosion susceptibility of soils adjacent to the proposed Ntabelanga dam (Parwada and van Tol, 2017), the second to determine the pollution from pit latrines to streams (Mamera and van Tol, 2018), and the third to quantify carbon stocks within the Ntabelanga dam footprint (van Tol et al., 2018). The soil observations were located in and around the proposed footprint. As with the Cathedral Peak samples, these samples are sufficient to evaluate the different disaggregation models, because they follow the catena specified in the LTS.

5.2.8 Model evaluation

All soil observations were performed completely independent of the LTS. Additionally, the small scale of samples enhances the actual evaluation as it captures the variability efficiently and large datasets are known for increasing the accuracy of identifying the predominant soil type. The number of samples used for evaluation is regarded as sufficient, as it compares well to the number of samples

used in other DSM projects in the area, such as 60 for land type disaggregation, 52 for an expert knowledge approach and 48 for a machine learning approach (van Zijl, 2019).

The soil observations from each catchment were used to evaluate each model based on their kappa statistics of the first most probable class raster. Additionally, the predictions confusion between the first and second most probable class rasters were also used to evaluate the models. The kappa gives an indication on the algorithms goodness of fit while confusion tells how certain the model is of its predictions. Therefore, these two indices were used as the main indicator of model performance. The confusion values were calculated according to Burrough et al. (1997) and described in (Chapter 3). Kappa was evaluated only on the first most probable class raster as this is a simplified model often necessary for decision making. The lower the confusion, the more certain the model is of the predictions. This measure is especially important where external soil observations are scarce, such as at Cathedral Peak.

5.3 Results and discussion

5.3.1 Overall model performance

The kappa and confusion for each algorithm can be seen in Table 5.5. Over both land types, NSC had a competitive kappa and confusion. Nearest shrunken centroid has an embedded feature selection which minimises unimportant covariate centroids to zero (Klassen, 2014; Tibshirani et al., 2003). It does so by shrinking the centroids to the average centroid for each class. Multinomial ridge regression had competitive kappa and confusion. However, the confusion was relatively higher at Cathedral Peak. The MLR algorithm penalises unimportant covariates based on the squared errors, thereby minimising the unimportant coefficients (Friedman et al., 2010). Therefore, these robust linear models prevent collinearity and overfitting which added to their performance.

Linear support vector machines had competitive kappa but had a comparatively higher confusion at Ntabelanga. Radial support vector machine performed well in terms of confusion, however, SVR had a relatively low kappa at Ntabelanga. Stochastic gradient boosting and RF also performed well in terms of confusion. This indicates that SVR, SGB, and RF although never the top performing algorithms, have little confusion between the first and second most probable soil associations across the land types. The original implementation through C5 performed well at Cathedral Peak in terms of kappa but performed poorly for all other performance indices.

Table 5.5: Algorithm performance showing kappa and confusion for Cathedral Peak and Ntabelanga.

Algorithms	Cathedral Peak		Ntabelanga	
	Kappa	Confusion	Kappa	Confusion
C5	0.40	0.60	0.08	0.89
KNN	0.23	0.46	0.07	0.92
LDA	0.27	0.53	0.12	0.93
MLP	0.11	0.49	0.08	0.85
MLR	0.34	0.55	0.17	0.84
NSC	0.43	0.22	0.11	0.85
RF	0.26	0.45	0.07	0.85
SGB	0.42	0.49	0.09	0.83
SVL	0.41	0.54	0.11	0.90
SVR	0.36	0.35	0.05	0.83

Although most complex algorithms performed well, they may be unnecessary due to an already computationally heavy method, lack of improved results, and a decrease in interpretability. Additionally, algorithms such as SGB, MLP, and SVR failed to classify three out of the five soil associations at Ntabelanga. Random forest, C5, and SVR failed to classify two of the soil associations. This can be attributed to these algorithms over classifying the pedo association during each realisation. The over classification perpetuates through the realisations resulting in a great under representation of the other soil associations. This was also seen by Holmes et al. (2015), who attributed this effect to the classes with low probabilities not being well correlated with the covariates. Here the effect is attributed to the algorithm used for each model as well as a weak correlation with covariates. This effect was observed after up-sampling the soil associations with the least percent area. However, without up-sampling, this trend was amplified, and performance dropped for all models. In contrast, linear models such NSC, MLR, LDA, and SVL only failed to classify the apedal association at Ntabelanga increasing their performance measures.

From these results, it is clear that some algorithms will perform better in either kappa or confusion values. Therefore, an algorithm should be chosen based on availability of observations to evaluate on. For example, if there are a few observations in an area, then a model that minimises the confusion might be more appropriate. This approach is more notable in areas with little knowledge of the soil distribution. If there are many observations, then the soil scientist can use expert knowledge to choose an algorithm based on specific needs or run many models and choose the best one. The latter is more suitable when disaggregating smaller areas such as a single land type. However, it was

found that evaluating one realisation was a good indicator of model performance. Therefore, in the case of large areas, one realisation can be evaluated and then the best model was used.

5.3.2 Cathedral Peak

Nearest shrunken centroid achieved the highest kappa (moderate agreement), however SGB, SVL, and C5 had similar values. Where the NSC algorithm stands out is in the model confusion, where it is substantially lower than the next algorithm (SVR). Therefore, NSC is considered the best model for Cathedral Peak. The NSC predictions and confusion at Cathedral Peak are shown in Figure 5.5. It should be noted that both C5 (76%) and KNN (73%) had a higher overall accuracy than NSC (71%), and SGB, SVL, and LDA achieved the same accuracy. Additionally, Botha (2016) achieved a kappa of 0.66 and an accuracy of 88% through an expert knowledge approach at Cathedral Peak. However, the approach presented in this study, is much more automated making it more suitable for larger areas (van Zijl, 2019) and is not limited to the TMU boundaries (Odgers et al., 2014).

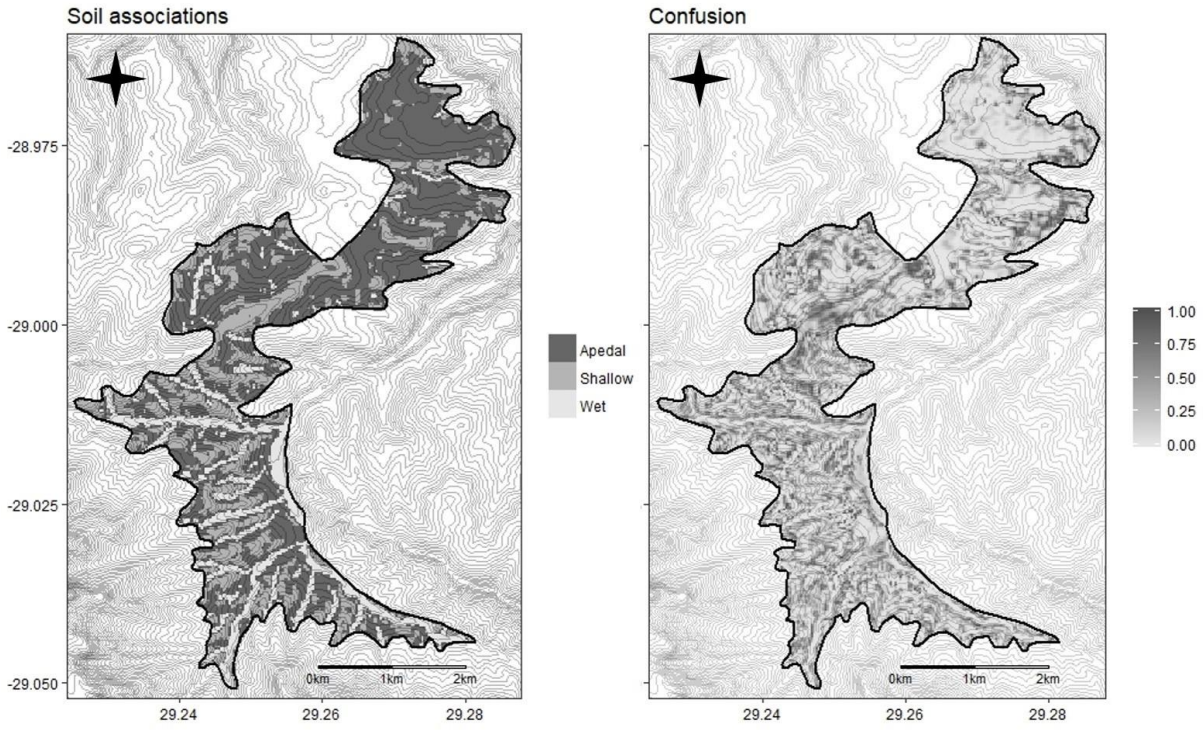


Figure 5.5: Nearest shrunken centroid predictions and confusion at Cathedral Peak.

For NSC, the confusion is highly correlated to soil association and TMUs. To compare, the confusion was analysed by a post hoc Tukey-Kramer ($P < 0.05$*) implemented through a residual maximum likelihood model (REML). The REML model is necessary to account for spatial auto-correlation and is

considered best practice with spatial data (Lark and Cullis, 2004). Wet soils had the lowest confusion of 0.18*, followed by apedal (0.21*), and shallow soils (0.25*). This is surprising, as wet soils were classified with less accuracy (50%) followed by that of apedal soils (68%). Shallow soils were classified with the highest accuracy of 85%which were found over both crest and mid-slope positions. Crests had the lowest confusion of 0.01* followed by valley (0.05*) and mid-slopes (0.25*). This indicates that soils found over many TMUs and TMUs with many soil associations will have the highest confusion.

At Cathedral Peak, most of the algorithms performed rather well. This was attributed to the clear toposequence and relatively non-complex soil pattern in the area. Shallow soils occur on crest positions where erosion exposes the lithic contact, apedal soils are found on mid-slopes where erosion is less, allowing more profile development, and wet soils occupy the valleys where water accumulates. These results largely confirm the LTS data and the study by Botha, (2016), as soils are strongly tied to the TMUs on the land type.

5.3.3 Ntabelanga

There is no clear algorithm which substantially outperformed another at Ntabelanga. However, MLR achieved the highest kappa (slight agreement) but had similar confusion to SGB and SVR. The MLR model predictions and confusion can be seen in Figure 5.6. The MLR model achieved an accuracy of 33% which is similar to that of Botha (2016), who reached a kappa of 0.20 and an accuracy of 37% at Ntabelanga. Additionally, these results are comparable to van Zijl et al. (2013), who achieved an accuracy of 35% when disaggregating two land types with five soil associations. However, MLR overclassified both shallow and wet soils.

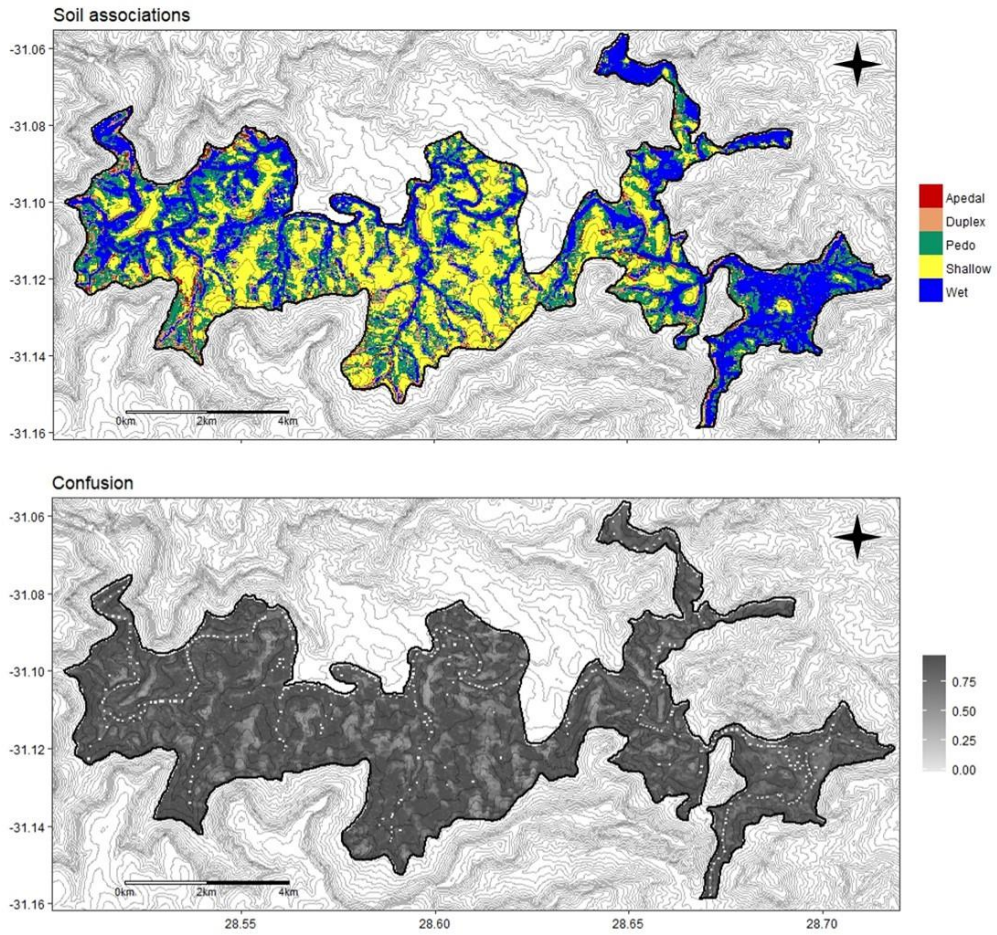


Figure 5.6: Multinomial ridge regression predictions and confusion at Ntabelanga.

The MLR model's confusion had a similar trend to the NSC model at Cathedral Peak. The confusion was lowest for apedal soils of 0.56* followed by shallow (0.76*), wet (0.85*), pedo (0.92*) and duplex soils (0.95*). Soil associations which did not show a trend over the TMUs had the highest confusion. For example, shallow soils had a relatively low confusion but are found on every TMU, however, there is a clear trend from mid-slope to valley. In contrast, pedo soils are found on every TMU but with no clear trend. Therefore, pedo soils have a relatively high confusion. Additionally, all soil associations are found on foot-slopes which also had the greatest confusion (0.89*) followed by valleys (0.86*) and mid-slope positions (0.76*). This was expected and confirms TMUs with the greatest amount of soil associations will have the highest uncertainty.

In general, the models did not perform as well in Ntabelanga as in Cathedral Peak. Soil distribution patterns in the area are governed chiefly by the geology/lithology and secondary by the topography. Low model performance due to complex geological relationships was also found by Holmes et al. (2015) and is due to the TMUs not aligning the samples in the correct feature space. The horizontal

and vertical variation within lithological layers e.g. sandstone, mudstone and dolerite within the Tarkastad and Beaufort groups, results in considerable variation in soils over short distances. This was clear during the field survey where it was observed that strongly structured soils (duplex and pedo) were found on mudstones and apedal soils were found on sandstones. Soils derived from dolerite, consisted of both red apedal and pedo soils. There is also colluvial material creating binary profiles which added to the complexity. In addition, the permeability of the different parent materials impact weathering patterns (depth) and soil/bedrock flow paths (wetness). This, together with the low relief, results in the occurrence of wet and shallow soils throughout the study area.

In an attempt to improve these results, model averaging of the overall top performing algorithms in terms of kappa (MLR, NSC, SVL, LDA, SGB) and model averaging using the algorithms which classified a particular soil association best (MLR, MLP, SGB, NSC), were tried. However, the results were disappointing only achieving a kappa of 0.05 and 0.06, respectively. Alternatively, if a geology/lithology map were available, expert knowledge could be used to determine the probability of each soil association on the different categories instead of using the LTS probabilities and TMUs. Additionally, expert rules could be used for soil type allocation during the resampling procedure as implemented by Häring et al. (2012) and Vincent et al. (2016). However, there was no reliable geology/lithology data available in an area to use as a covariate to train on. Although, the attempt failed, this shows the versatility of such the modified DSMART approach.

Although results were poor at Ntabelanga, it shows some possibilities the modified DSMART model can perform. For example, running many algorithms and choosing the best one or using model averaging. It also provides additional functions that can be used to optimise algorithm parameters such as cross-validation at each realisation. Additionally, covariates can be pre-processed allowing for Box Cox and principle component analysis transformation. These additional techniques can be further explored when disaggregating a particular region.

Internationally, DSMART disaggregation produced accuracies comparable to other disaggregation studies. For example Møller *et al.* (2019), notes that 17 – 23% accuracy is the range when evaluating disaggregation approaches for the first most probable class raster when disaggregating over a large area with many soil types. Even with the insufficient results achieved at Ntabelanga, the accuracy was far above the international norm. This can be attributed to soil form aggregation and the scale of the

site. Additionally, other research such as Møller et al. (2019), comes to the conclusion that a more detailed soil map is better than a less accurate soil map. When aggregating soils, it seems accuracy is more important than a detailed map in South African conditions. This is more important for land use management as the scale used, might not be appropriate for environmental modelling.

5.3.4 Covariate importance

The five most important covariates for NSC at Cathedral Peak and MLR at Ntabelanga are shown in Table 5.6. The covariate importance for NSC is calculated as the difference between a particular shrunken centroid for a class and that of the overall centroid (Tibshirani et al., 2003). Therefore, the larger the difference, the more important that covariate. The covariate importance for MLR is calculated on the absolute value of the coefficients. The further away from zero, the more important that covariate is. Both importance measures were averaged over all realisations and scaled to percentages.

Table 5.6: Five most important covariates and their descriptive statistics for the NSC and MLR algorithms at Cathedral Peak and Ntabelanga, respectively.

Cathedral Peak	Covariate	Mean (%)	Sd (%)	Lower (%)	Upper (%)
1	TPI	64	39	59	68
2	DC	62	32	59	66
3	LC	59	37	55	64
4	Convexity	43	26	40	46
5	SWI	39	23	37	42
Ntabelanga					
1	Aspect	36	27	34	38
2	Terrain factor	35	25	33	37
3	DEM	34	25	32	37
4	Convexity	31	23	29	33
5	NO	31	22	29	33

At both sites, no covariate was the most important for all realisations. Therefore, the models are utilising different covariates for each realisation. This could indicate that the models' need many covariates to capture the soil distribution in the two areas. It was thought that this could also be a function of the number of samples per realisation. However, when performing the models with 50 samples per TMU, there was no difference in either model performance or covariate importance.

Therefore, it could indicate that the covariates were collinear, and the number could have been reduced at Cathedral Peak, or that the covariates were not sufficient in the case of Ntabelanga.

Covariates which characterise slope position, slope shape, and water accumulation are the most important covariates at Cathedral Peak. This is no surprise as the soils are clearly controlled by landforms in the area. Apedal soils are correlated most with catchment slope as apedal soils are found on sloping positions. Both wet and shallow soils are strongly correlated with TPI as wet soils are in valley positions and shallow soils are on crest positions. However, the overall importance of catchment slope was low and TPI importance varied greatly with each realisation.

Covariates which characterise sun angle and amount, elevation, and slope shape correlate most with the soil associations at Ntabelanga. This is surprising as no spectral covariates were characterised even in the 10 most important covariates. It was thought that the spectral covariates would give more insight into the soil distribution, however, this was not clearly shown in these models. This was also seen by Møller et al. (2019), who found Landsat 8 bands and vegetative indices had a low importance. This could be due to the target-based soil assignment on TMUs which focuses more on topographic relationships. However, high soil erosion might have induced these results making the spectral covariates ineffective.

5.4 Conclusion

A modified DSMART model was developed to test 10 algorithms on their ability to disaggregate the LTS. The algorithms were compared on two environmentally contrasting land types in South Africa. The land types were first stratified with TMUs and these TMUs were used for re-sampling in DSMART. The algorithms were evaluated on the kappa of the first most probable class raster and the confusion between the first and second most probable class raster. The main findings of this study are:

- Robust linear algorithms such as NSC and MLR, were the top performing models for Cathedral Peak and Ntabelanga, respectively.
- When disaggregating a single land type, complex models do not improve the results and are less computationally efficient.
- Where there are strong soil-terrain relationships, the method produced satisfactory results such as Cathedral Peak.

- Where there are strong soil-geological relationships, the method was deemed unfit such as in Ntablenga. Alternatively, another input map could be tried which does not focus on TMUS and relies more heavily on parent material.
- Grouping soil classes may be necessary when disaggregating soil maps with no legacy point data.
- Model averaging did not improve the results in the area with strong soil-geological relationships indicating the need to be supplemented with geological/lithological information.
- The results achieved, were comparable to other LTS disaggregation methods such as expert knowledge. However, this method is more automated making it more cost effective.

This study highlights the versatility the modified DSMART model brings to disaggregating the LTS. The modified DSMART allows users to choose the algorithm based on expert knowledge of an area, run many models to determine the best model, the ability to use model averaging, and/or optimise algorithm parameters. This methodology has implications for international datasets such as SOTER which also heavily relies on terrain to determine the soil distribution and which covers much of Southern Africa. This should be a priority in further research.

5.5 References

Bannari, A., Morin, D., Bonn, F., Huete, A.R., 1995. A review of vegetation indices. Remote Sens. Rev. 13, 95–120. https://doi.org/10.1080/02757259509532298

Botha, C.C., 2016. Disaggregating of land type data to aquire functional soil information (MSc thesis). University of the Free State.

Breiman, L., 2001. Random Forests. Berkeley, California. https://doi.org/10.1017/CBO9781107415324.004

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth Int. Group, CA.

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

Buhlmann, P., Yu, B., 2002. Analyzing bagging. Ann. Stat. 30, 927–961.

Bui, E.N., Loughhead, A., Corner, R., 1999. Extracting soil-landscape rules from previous soil surveys. Aust. J. Soil Res. 37, 495–508. https://doi.org/10.1071/S98047

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

Burrough, P.A., Van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: Spatial correlation, confusion and boundaries. Geoderma 77, 115–135. https://doi.org/10.1016/S0016-7061(97)00018-9

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analysis (SAGA). Geoscientific Model Development. https://doi.org/doi:10.5194/gmd-8-1991-2015

Dijshoon, J., van Engelen, V., JRM, H., 2008. Soil and landform properties for LADA partner countries (Argentina, China, Cuba, Senegal, South Africa and Tunisia).

Dixon, J., Gulliver, A., Gibbon, D., 2001. Farming Systems and Poverty: Improving Farmers' LiveliHoods in a Changing World. FAO and World Bank, Rome and Washington DC.

Evans, I.S., 2012. Geomorphometry and landform mapping: What is a landform? Geomorphology 137, 94–106. https://doi.org/10.1016/j.geomorph.2010.09.029

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1–24. https://doi.org/10.18637/jss.v033.i01

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207. https://doi.org/10.1016/j.geoderma.2009.06.003

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma 185–186, 37–47. https://doi.org/10.1016/j.geoderma.2012.04.001

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. Springer Series in Statistics.

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

Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. CSIRO 53, 865–880.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw- Hill, NY. https://doi.org/10.2307/211491

Klassen, M., Kim, N., 2009. Nearest Shrunken Centroid as Feature Selection of Microarray Data, in: Proceedings of the ISCA 24th International Conference on Computers and Their Applications. New Orleans, Louisiana, USA.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, the R.C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., 2018. Caret: Classification and Regression Training.

Land Type Survey Staff, 1972-2006. Land Types of South Africa on 1:250 000 scale. Pretoria, South Africa.

Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. Eur. J. Soil Sci. 55, 799–813. https://doi.org/10.1111/j.1365-2389.2004.00637.x

Mamera, M., Van Tol, J.J., 2018. Application of Hydropedological Information to Conceptualize Pollution Migration From Dry Sanitation Systems in the Ntabelanga Catchment Area , South Africa. Air, Soil Water Res. 11, 1–12. https://doi.org/10.1177/1178622118795485

McBratney, A.B., Santos, M.L.M., Minasnya, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Møller, A.B., Malone, B., Odgers, N.P., Beucher, A., Vangsø, B., Humlekrog, M., Minasny, B., 2019. Improved disaggregation of conventional soil maps. Geoderma 341, 148–160. https://doi.org/10.1016/j.geoderma.2019.01.038

Nauman, T.W., Thompson, J.A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385–399. https://doi.org/10.1016/j.geoderma.2013.08.024

Nauman, T.W., Thompson, J.A., Rasmussen, C., 2014. Semi-Automated Disaggregation of a Conventional Soil Map Using Knowledge Driven Data Mining and Random Forests in the Sonoran Desert, USA. Photogramm. Eng. Remote Sens. 80, 353–366. https://doi.org/10.14358/PERS.80.4.353

Odgers, N., Malone, B.P., 2017. rdsmart: Disaggregation and harmonisation of soil map units through resampled classification trees (R package version 2.0.3).

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100. https://doi.org/10.1016/j.geoderma.2013.09.024

Parwada, C., Van Tol, J., 2017. Soil properties influencing erodibility of soils in the Ntabelanga area, Eastern Cape Province, South Africa. Acta Agric. Scand. Sect. B - Soil Plant Sci. 67, 67–76.

R Core Team, 2017. R: A language and environment for statistical computing.

Ray, S.S., Singh, J.P., Das, G., Panigrahy, S., 2004. Use of high resolution remote sensing data for generating site-specific soil mangement plan, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Schulze, R.E., 2007. South African Atlas of Climatology and Agrohydrology, WRC Report 1489/1/06, Section 1.3. Water Research Commission, Pretoria, South Africa.

Silva, S.H.G., Menezes, M.D. de, Owens, P.R., Curi, N., 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma 267, 65–77. https://doi.org/10.1016/j.geoderma.2015.12.025

Soil Classification Working Group, 1991. Soil Classificatioin: a Taxonomic System for South Africa, 2nd ed. Department of Agricultural Development, Pretoria, South Africa.

Strobl, C., Malley, J., Tutz, G., 2009. Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychol. Methods 14, 323–348. https://doi.org/10.1037/a0016973

Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334–345. https://doi.org/10.1016/j.geoderma.2013.08.018

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2003. Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. Stat. Sci. 18, 104–117. https://doi.org/10.1163/15718093-12341267

van Tol, J.J., Akpan, W., Kanuka, G., Ngesi, S., Lange, D., 2014. Soil erosion and dam dividends: science facts and rural 'fiction' around the Ntabelanga dam, Eastern Cape, South Africa. South African Geogr. J. 98, 169–181. https://doi.org/10.1080/03736245.2014.977814

van Tol, J.J., Akpan, W., Maroyi, A., Mutengwende, N., Huchermeyer, N., Ngesi, S., Nqandeka, H.M., Mamera, M., Bradley, G., Rowntree, K.M., 2018. The Mzimvubu Water Project: Baseline indicators for long-term impact monitoring. WRC Proj. No. K5/2433.

van Zijl, G., 2019. Digital soil mapping approaches to address real world problems in southern Africa. Geoderma 337, 1301–1308. https://doi.org/10.1016/j.geoderma.2018.07.052

van Zijl, G.M., Botha, J.O., 2016. In pursuit of a South African national soil database: potential and pitfalls of combining different soil data sets. South African J. Plant Soil 1–8.

van Zijl, G.M., Le Roux, P.A., Turner, D.P., 2013. Disaggregation of land types using terrain analysis, expert knowledge and GIS methods. South African J. Plant Soil 30, 123–129. https://doi.org/10.1080/02571862.2013.806679

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2016. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142. https://doi.org/10.1016/j.geoderma.2016.06.006

Weiss, A.D., 2000. Topographic Position and Landforms Analysis, in: ESRI User Conference.

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

Chapter 6 Input map and feature selection for soil legacy data

This chapter is based on (Under review) Flynn, T., Rozanov, A., Clarke, C., Input map and feature selection for soil legacy data. *Geoderma*.

Abstract:

Soil legacy disaggregation techniques are becoming more relevant, as cost effective highly detailed soil information is required to advise agriculture, hydrological, ecological, engineering, and a variety of other disciplines. Disaggregation involves the spatial prediction of individual soil classes from soil legacy polygons which have multiple soil classes, while specifying the approximate proportion of each soil class and verbally or diagrammatically explaining their distribution in the landscape. However, DSMART is computationally intensive and has many parameters that must be optimised. This study aimed to address these drawbacks including input map selection, feature selection, and resample size optimisation. The research site was selected in the upper reaches of the Mvoti river catchment covering 317 km^2 in KwaZulu Natal province, South Africa. The catchment consists of 20 soil-terrain polygons drawn at a 1:250,000 scale from the South African Land Type Survey (LTS). First, the optimal input map derived from landform elements (geomorphons) was selected through a spatially resampled Cramer's V test to determine the association between the legacy polygons (proportion of terrain) and the geomorphon units. This was done for five different aggregated geomorphons with different resolutions and parameters. Second, three feature selection algorithms (FSAs) were embedded into DSMART to determine if the algorithms could improve accuracy and computationally efficiency. Third, the FSAs were compared using 25, 50, 100, and 200 resamples per polygon. The results indicate that the Cramer's V test is a rapid method to determine the optimal input map. All FSAs achieved a significantly greater accuracy then when disaggregating the original legacy polygons and were more computationally efficient than when using all 52 covariates. This study has implications when disaggregating large and small datasets by improving computational efficiency while maintaining an acceptable accuracy.

6.1 Introduction

Disaggregation of soil legacy data is a relatively recent technique; and can increase the detail of existing soil maps with few resources (Odgers et al., 2014). Disaggregation uses soil polygons with an associated legend to spatially predict individual soil classes (McBratney, 1998b). This is particularly important in southern Africa where resource allocation to new soil surveys is limited, yet agriculture and environmental concerns continue to grow (Ranst et al., 2010). In general, South Africa has

focused on disaggregating a single soil-terrain polygon through an expert driven approach (van Zijl, 2019). However, advancements in available disaggregation algorithms such as DSMART allow disaggregating multiple soil-terrain polygons simultaneously.

Advances in the DSMART method have the benefit of an increase in accuracy such as area proportional sampling (Møller et al., 2019; Vincent et al., 2016) and a reduction of computation time such as producing multiple realisations from a single set of resamples (Chaney et al., 2016; Møller et al., 2019). Additionally, any machine learning algorithm may be implemented within DSMART as demonstrated in (Chapter 5), making it adaptable in terms of soil complexity, location, and scale. However, Zeraatpisheh et al. (2019), notes that DSMART is computationally demanding and requires specific model parameters such as the number of realisations, resample size, and algorithm selection. Furthermore, Holmes et al. (2015) and Møller et al. (2019), state that the prediction accuracy heavily relies on which input map is used for resampling. Therefore, there is a necessity to improve the methodology to increase accuracy and improve computational efficiency.

Past studies have incorporated soil-landscape rules either through allocating soil classes to specific covariate values (Vincent et al., 2016) or by manipulating the input maps to incorporate these rules during the resampling procedure as performed by Møller et al. (2019) and shown in Chapter 4 and 5. However, when using the latter approach, many input maps can be developed or are available, and it is unknown which input map will produce the highest accuracy. This makes deciding which input map to resample, either a time-consuming process as it is unclear which is best until DSMART predictions are evaluated or subjective relying on expert opinion.

Previous studies into DSMART have relied on either a large pool of covariates (Møller et al., 2019) or expert selection of covariates as done in Chapter 4 and 5. The former approach relies on the embedded feature selection of machine learning algorithms such as decision trees and RF. However, decision tree performance is known to decrease when unimportant covariates are introduced to the model (Kohavi and John, 1997). Furthermore, because of the random selection of covariates at each split of a decision tree, unimportant covariates might be selected and accuracy can decrease in RF models (Kuhn and Johnson, 2013). Nevertheless, it is often difficult to determine which covariates will correlate best with soil classes and as both Vincent et al. (2016) and Odgers et al. (2014) have stated, selecting the appropriate covariates should improve prediction accuracy.

Considered an essential part of any machine learning model, FSAs are one option to quantitatively select covariates. Feature selection can be done by finding a set of transformed covariates that represent the most environmental variability (unsupervised) or finding a set of covariates which correlate best to the soil classes (supervised). This is important not only to improve prediction accuracy and computational efficiency, but also to increase model interpretability (Guyon and Elisseeff, 2003) and produce more stable results (Larose, 2006). However, when using soil polygons to resample, soil classes have a spatially uncertain location which may lead to important covariates not correlating with soil classes. Therefore, supervised feature selection is difficult when there are no georeferenced point observations for the FSAs to select covariates from.

Resample size is another possible limitation to the computational efficiency and accuracy of DSMART. Machine learning is a data driven approach, where in general, the more training data there is, the higher accuracy of the model (Hastie et al., 2009). Møller et al. (2019), states that resample size can be increased to improve prediction accuracy when using RF. However, an optimal resample size has yet to be explored and the optimal number may have a large computational cost. The objective of this study (Framework 2, Objective 5) was to address these three limitations of DSMART by introducing a rapid input map selection technique, incorporating FSAs, and determining appropriate resample sizes.

6.2 Materials and Methods

6.2.1 Research site

The study site was the Mvoti catchment (Figure 6.1 in Kwa-Zulu Natal Midlands, South Africa (30° 19′ 4.4″ E to 30° 38′ 23″ E and 29° 16′ 36.5″ S to 29° 5′ 22.6″ S). The catchment is 317 km^2 and ranges in altitude from 950 m to 1540 m above sea level. The catchment has an Ustic soil moisture regime with precipitation ranging from 800 mm/year at low altitudes to 1600 mm/year at higher altitudes (Wiese et al., 2016). The geology of the western region of the catchment is characterised by shale of the Pietermaritzburg formation of the Ecca group. Dolerite dykes are common which appear sporadically among the Karoo sediments often forming rather large features particularly at the foothills of the mountain ranges (Land Type Survey Staff, 1972 - 2006). Mvoti vlei a large wetland covers 29 km^2 (9%) predominantly in the eastern lower reaches of the catchment (Nel et al., 2011). Crop production

includes maize and sugar cane which are grown at the foothills while pine and eucalyptus plantations are found at the mountain slopes along the grasslands (Wiese et al., 2016).

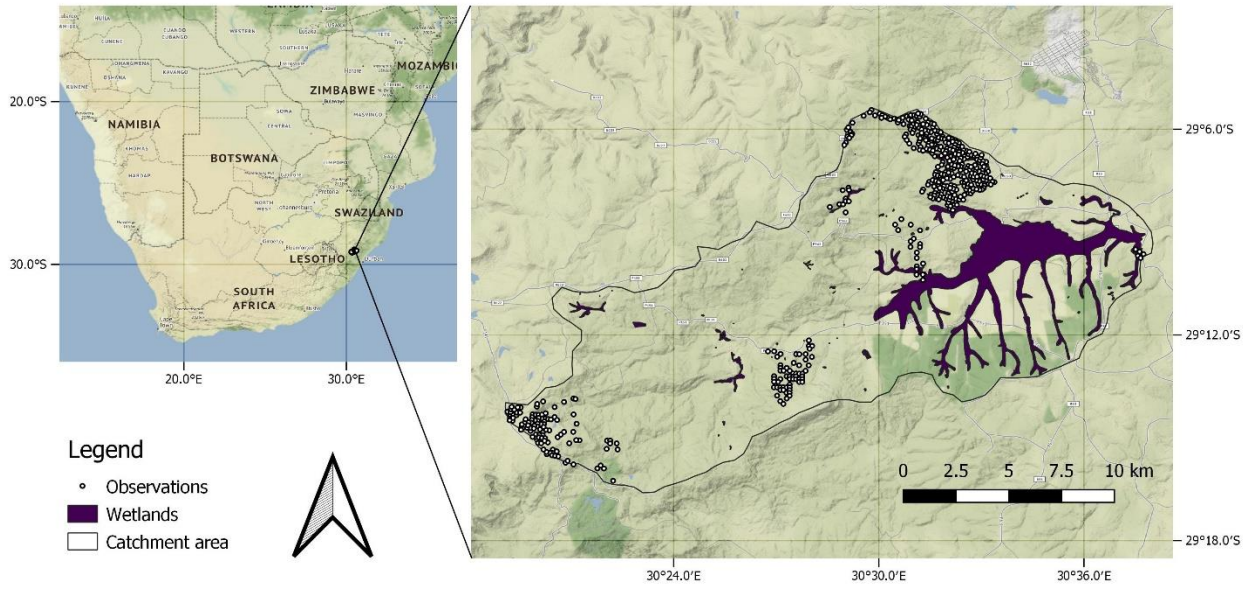


Figure 6.1: Mvoti catchment within South Africa, 500 soil observation, and the wetlands in the catchment (Stamen terrain map).

6.2.2 Soil legacy data

The legacy soil-terrain information was obtained from the LTS (Land Type Survey Staff, 1972 - 2006). The catchment contains 20 land types which range in size from 1 ha to 4,900 ha shown in Figure 6.2. It should be noted that some land types extend beyond the catchment boundaries and therefore, are small in area within the catchment.

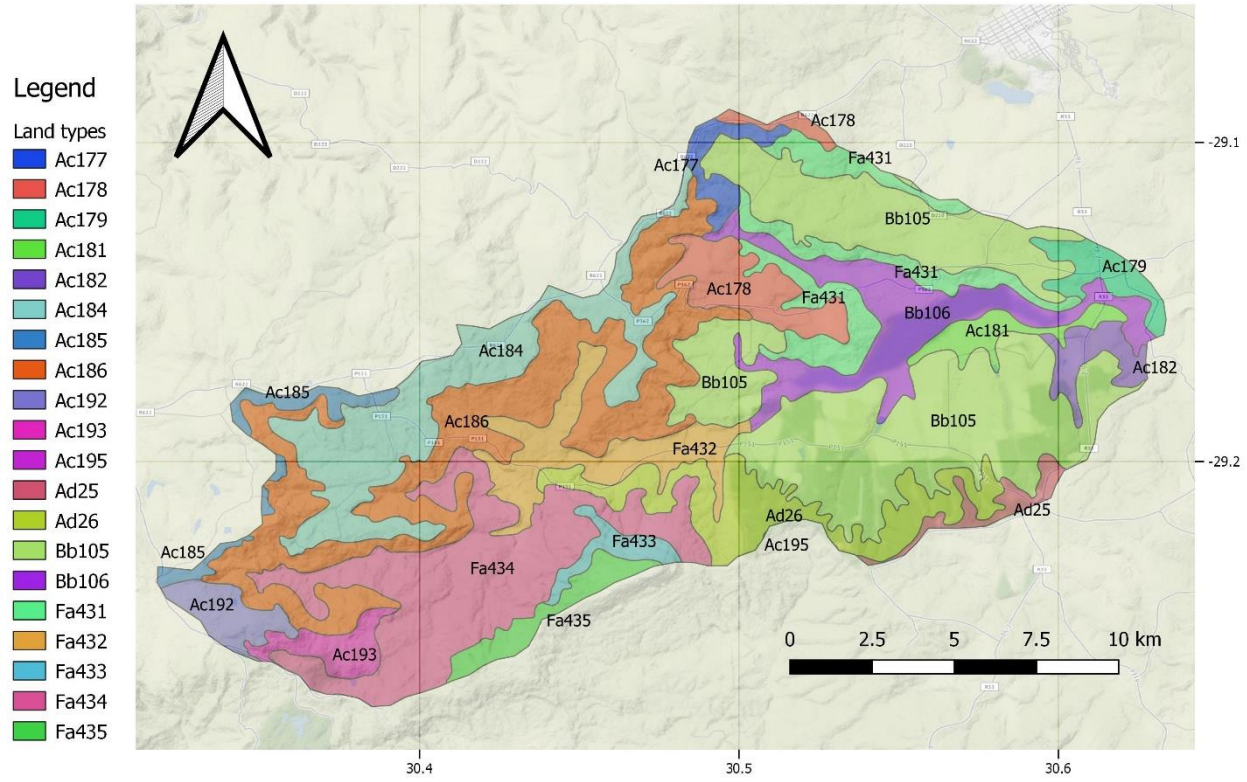


Figure 6.2: The 20 LTS polygons which fall within the Mvoti catchment.

There are 26 South African soil forms (Soil Classification Working Group, 1991) specified in the 20 land types of the study area. These soil forms were aggregated into soil associations based on morphological properties which is a common procedure when using South African soil taxonomy. The soil associations, their USDA Great Group (Soil Survey Staff, 2014) equivalence, and description are shown in Table 6.1.

The aquic association was based on a shallow water table with hydromorphic features. The cambic association was based on Inceptisols with an apedal structure without lithic contact. Chromic soils are highly weathered soils rich in Fe and/or Mn with or without clay illuviation. The leptic association is the same as the cambic association with lithic contact. Luvic soils have a strongly structured B horizon and comprise of Alfisols. It should be noted that neither South African soil forms or Great Groups were used for predictions as they differed between the LTS data and evaluation data. Additionally, it is difficult to translate soil forms to Great Groups leaving room for misinterpretation.

Table 6.1: Soil associations, Great Group equivalent, and soil description found in the Mvoti catchment.

Association	Great groups	Description
Aquic	Endoaquepts Endoaqualfs	Soils with a water table < 25 cm from surface
Cambic	Typic/Aquic Haplustepts	Slightly weathered soils with apedal subsoil
Chromic	Haplustox Haplustults Rhodustults	Highly weathered soils with or without clay illuviation
Leptic	Lithic Haplustepts	Soils with lithic contact below epipedon
Luvic	Haplustalfs Kandiustalfs	Soils with a strong subsoil horizon < 50 cm

6.2.3 Processes

The processes used in this study is shown in Figure 6.3. This involved creating multiple input maps through geomorphons to disaggregate into soil associations. Therefore, the original LTS polygons were not used for disaggregation. The best input map was quantitatively selected through the association of TMUs in the LTS legend and aggregated geomorphons through a resampled Cramer's V test (CV). DSMART was then run with three FSAs with different resample sizes on the best fitting input map. Each model was run with 10 realisations and was evaluated on external soil observations. Feature selection algorithms were also evaluated on their relative efficiency in terms of accuracy, runtime, and number of covariates selected through a relative efficiency index. The model which showed an adequate accuracy with an acceptable efficiency was taken as the final model.

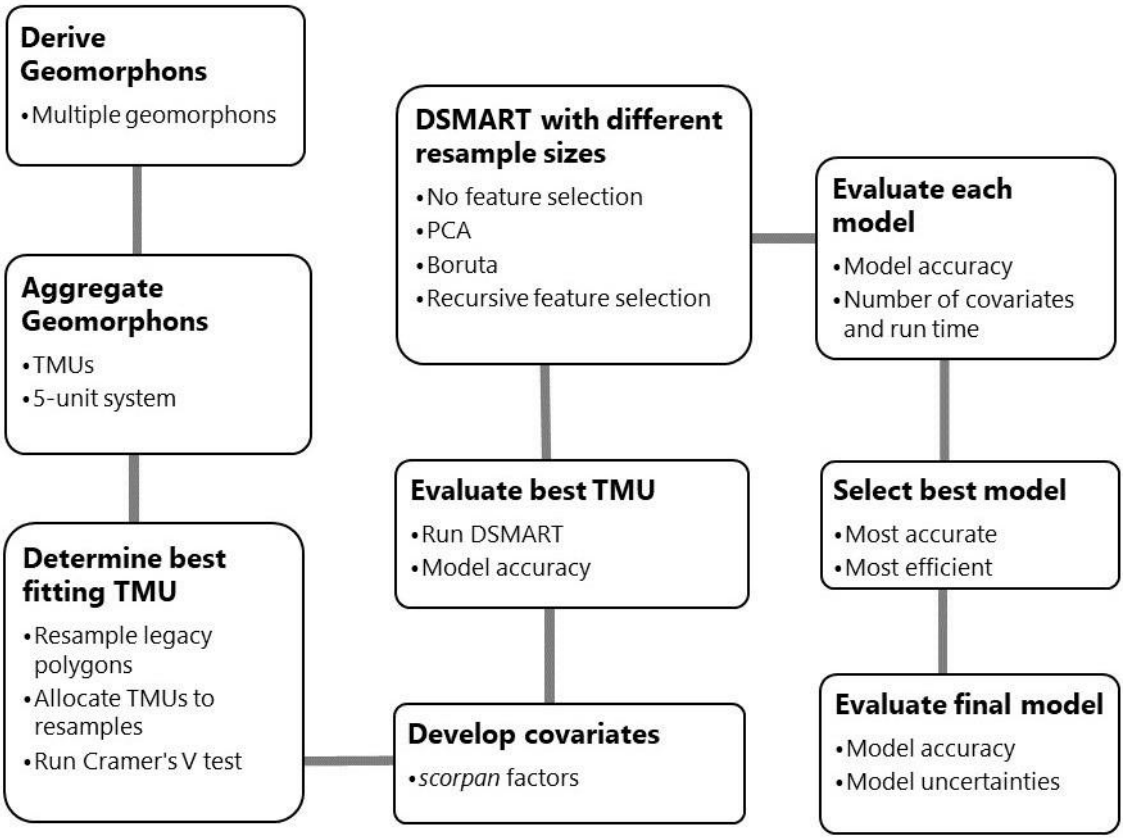


Figure 6.3: Flow chart of process used to evaluate input maps, feature selection algorithms, and production of final model.

6.2.4 Input maps

The catchment was stratified by geomorphons (Jasiewicz and Stepinski, 2013)to obtain TMUs. Five geomorphon sets were derived at a 30 m resolution with search radii of 10 (GM-10), 25 (GM-25), 50 (GM-50), 100 (GM-100), and 200 (GM-200) cells (inner search radius was kept at zero). The geomorphons were developed through the r.geomorphons add-on in the Geographic Resources Analysis Support System (GRASS Development Team, 2017). The selection of a 30 m resolution roughly corresponds to the 1:50,000 topographic sheets the TMUs were derived from. The five sets of geomorphons were aggregated into a 5-unit system to correspond with each land types TMU. The aggregation approach follows the method developed in Chapter 4. Notably, scarp was not classified during the aggregation method and therefore, only 4 TMUs are possibly classified.

Only the TMUs were used to disaggregate the catchment into soil associations allocated through the soil probabilities in the LTS legend. This approach limits the soil allocation procedure to only the TMUs and therefore, the original LTS polygons were not used for disaggregation. This approach

differs from that of Møller et al. (2019) and Chapter 4. This method was chosen as it was seen that the soil associations were found throughout the catchment. Therefore, the LTS polygons were deemed unnecessary and if used, would decrease the computational efficiency as more polygons would have been disaggregated. For example, TMUs consist of four polygons, the LTS consists of 20 polygons, and combining both the TMU and LTS would produce 56 polygons.

6.2.5 Input map evaluation

To determine the best fitting input map through the LTS information, a resampled CV test was utilized. This process is similar to the resampling and soil class allocation procedures of DSMART. One-hundred random samples were taken from the original LTS polygons, allocated a TMU according to the probabilities in the LTS legend, and a CV test was run against all sets of aggregated geomorphons. One-thousand iterations were used to get the average CV value and confidence intervals. The aggregated geomorphon with the highest average CV was taken as the best fitting input map. Therefore, this map was used to estimate TMUs.

6.2.6 Model training

All models were developed in R code (R Core Team, 2017) and machine learning parameters optimised in the caret R package (Kuhn et al., 2018). For each experiment, the catchment was disaggregated through a modified DSMART model run with 10 realisations which is equivalent to 1000 realisations in the original implementation of DSMART. Instead of the original C.5 decision trees (Quinlan, 1993), RF (Breiman, 2001) was used as implemented in the randomForest R package (Liaw and Wiener, 2002).

Random Forest was implemented in DSMART because it is resistant to over-fitting, can handle non-linear relationships between soil classes and covariates (Grunwald et al., 2011), does not assume any soil class distribution (Hengl et al., 2015), and has been shown to be computationally efficient (Møller et al., 2019). The primary reason why RF is more efficient is that only one set of resamples is needed to produce *n* number of realisations through bagging the decision trees. The final prediction is the modal soil association for the ensemble of decision trees at a given pixel (Breiman, 2001).

The RF models were implemented with 100 trees (*ntree*) and the *mtry* parameter was set to the default where *mtry* is $\sqrt{number\ of\ covariates}$ (Liaw and Wiener, 2002). Growing 100 trees is

analogous to running 100 realisations in the original implementation of DSMART. However, instead of resampling the polygons for each realisation, resamples are performed through bootstrapping the samples of each decision tree ("bagging"; Breiman, 1996). This is similar to the method of Chaney et al. (2016) and Møller et al. (2019) where one set of samples is taken from which, predictions and uncertainties can be estimated.

DSMART was run with 25, 50, 100, and 200 resamples per polygon for three feature selection algorithms (FSAs) as well as run with all covariates. Additionally, DSMART was run on the original LTS polygons using all covariates (control). Area proportional sampling was not used as too keep the number of samples the same across the polygons. This was important to compare the number of resamples run.

6.2.7 Model evaluation

Nine hundred soil observations were obtained from three different survey projects from Mondi Forests Ltd conducted at a scale of a detailed survey (1:25,000) and a confusion index between the first and second most probable class rasters as described in Chapter 4. The soil observations were placed in a systematically aligned grid which was done completely independent of the LTS. This dataset was used because it represents the largest number of observations in the catchment.

The number of soil observations was reduced to 500 (Figure 6.1) through cLHS (Minasny and McBratney, 2006). Conditioned Latin Hypercube sampling was implemented to reduce the bias in the original observations and select a distribution of observations which represent a large degree of environmental variability. The cLHS was implemented on aspect, altitude, profile curvature, plan curvature, slope, and a SAGA wetness index (SWI) with 10,000 iterations (default setting). The South African soil classes, number of observations, USDA Sub Group equivalence and the soil associations classified in the evaluation dataset are shown in Table 6.2.

Table 6.2: Soil evaluation dataset showing South African soil forms, their frequency, Soil Taxonomy equivalence, and their soil association.

SA Soil Type	Frequency	Sub Group	Association
Bloemdal (Bd)	3	Aqueptic Haplustox	Chromic
Clovelly (Cv)	35	Lithic Haplustox	Chromic
Griffin (Gf)	23	Typic Haplustox	Chromic
Glenrosa (Gs)	14	Lithic Haplustept	Leptic
Hutton (Hu)	57	Typic Haplustox	Chromic
Inanda (Ia)	145	Humic Rhodic Haplustox	Chromic
Katspruit (Ka)	4	Typic Endoaquent	Aquic
Kranskop (Kp)	43	Humic Haplustox	Chromic
Magwa (Ma)	81	Humic Xanthic Haplustox	Chromic
Mispah (Ms)	5	Lithic Haplustept	Leptic
Nomanci (No)	36	Lithic Humlustept	Leptic
Oakleaf (Oa)	8	Typic Haplustept	Cambic
Pinedene (Pn)	4	Oxyaquic Haplustox	Chromic
Shortlands (Sd)	1	Typic Rhodustults	Chromic
Sepane (Se)	2	Aquic Kandiustalfs	Luvic
Sweetwater (Sr)	9	Humic Haplustepts	Cambic
Swartlands (Sw)	2	Typic Kandiustalfs	Luvic
Tukulu (Tu)	1	Aquic Haplustepts	Cambic
Valsrivier (Va)	6	Typic Kandiustalfs	Luvic
Westleigh (We)	3	Plinthaquic Haplustox	Chromic

6.2.8 Feature selection

One unsupervised and two supervised FSAs were embedded into the DSMART model to determine if feature selection can increase accuracy and decrease computational power. The FSAs implemented

include principle component analysis (PCA) and two RF wrappers Boruta (Kursa and Rudnicki, 2010) and RFS.

Boruta tries to find all relevant covariates to the target soil classes. Boruta does this by selecting random covariates to create a "shadow" covariate to compare the other covariates against. Relevant covariates are selected based on a statistical test of independence. Boruta was used because it is a novel technique, computationally efficient, and has a few parameters to optimise. Recursive feature selection was implemented as described in Chapter 2. Recursive feature selection was implemented because it has been shown to improve predictions (Brungard et al., 2015).

Principle component analysis was conducted on all covariates during the initiation of DSMART. Therefore, PCA was only run once before running DSMART for each resample size. This created a transformed covariate set where the first components which captured 85% of the variation were selected. On the other hand, both supervised FSAs were run for each realisation in DSMART. This was necessary to relate the soil associations with the covariates. The supervised FSAs were run with their default settings to limit parameter optimisation. Boruta was run with 10 iterations, 500 trees gown, and with Bonferroni multiple class correction as the statistical test of independence. The RFS was run with 10 iterations and 500 trees grown for 52, 16, 8, and 4 randomly selected covariates. It should be noted that the FSAs do not produce realisations in DSMART, they only attempt to select the appropriate covariates. Therefore, the trees grown for each supervised FSA were increased to 500 to find an optimal subset of covariates.

6.2.9 Feature selection and resample size evaluation

The FSAs were evaluated for each resample size on their overall accuracy using the external soil observations. Additionally, the FSAs were evaluated on the number of covariates selected and runtime of the models. This evaluation is important as reducing the number of covariates increases model interpretability and decreases computational power. This was done by developing a relative efficiency index (REI) shown in Equation 6.1, where EI_{mean} is the average efficiency and EI_{max} is the maximum efficiency possible.

$$\mathrm{REI} = \frac{EI_{mean}}{EI_{max}} \tag{6.1}$$

The equation for EI_{mean} is shown in Equation 6.2, where *A* is the overall accuracy (or kappa*100), *C* is the number of covariates selected, and *T* is the average runtime of the algorithm in any time unit (depending on the runtime length). However, the units of *T* will affect how much emphasis the REI puts on *C*. For example, using minutes instead of hours will put a larger emphasis on runtime. When using hours instead of minutes, a larger emphasis will be placed on the number of covariates. Additionally, weights can be added to put more emphasis on either the number of covariates or runtime. Alternatively, *C* and *T* could be scaled from the minimum to maximum of *A* to put equal weights on all parameters.

$$\mathrm{EI_{mean}} = \frac{\mathrm{A}}{\mathrm{C+T}} \tag{6.2}$$

EI_{max} is a theoretical value which is calculated similar to that of EI_{mean}. However, *A* is the highest accuracy achieved, *C* is the lowest number of covariates selected, and *T* is the fastest run time of any model. Essentially, it is the highest efficiency theoretically possible. Alternatively, the highest EI_{mean} could be used as EI_{max} and therefore, the highest REI will be equal to one. The variable *T* units and weights should be the same as EI_{mean}. In this study, minutes were used with no weights and *C* and *T* were scaled.

6.2.10 Covariates

Fifty-two covariates were developed which include soil, relief, vegetation, parent material, and neighbourhood according to the *scorpan* factors (McBratney et al., 2003). One predictive model used all 52 covariates which relies on the embedded feature selection of RF. Additionally, the large pool of covariates were used to evaluate how well the FSA reduce data dimensionality before training the predictive models. The covariates developed are shown in Table 6.3. It should be noted that covariates that are known to be correlated were not removed to evaluate how well the FSA reduce these covariates.

Table 6.3: Covariates developed and their description.

	Covariates	Description
Relief	Aspect	Degrees from north
	Analytical Hillshading	Radiation direction from surface (0-360)
	Catchment area	Flow accumulation (upslope cells)
	Catchment slope	Average slope perpendicular to contours
	Convergence Index	Convergent and divergent areas
	Convexity	Amount of convexity
	Digital elevation (DEM)	Altitude from sea level
	Flow direction	Direction (aspect) of water movement
	Geomorphon	Landform elements (200 cell radii)
	Gradient	Downslope controls on local drainage
	Gradient difference	Difference local gradient
	LS factor	Slope length factor
	Mass balance index (MBI)	Landscape stability
	Mid-slope position	Height of middle slope
	MRTTF	Multiresolution ridge top flatness
	MRVBF	Multiresolution valley bottom flatness
	Negative openness	Enclosed landscape
	Normalized height	Normalized height from local minima
	Plan curvature	Horizontal curvature
	Positive openness	Open landscape
	Profile curvature	Vertical curvature
	Sky view	Index of visible sky
	Slope	Scalar hill steepness (Degrees)
	Slope height	Height of slope from local minima
	SWI	SAGA wetness index
	Terrain factor	Shading dependent of light source
	TPI	Terrain position index
	Terrain_view	Shading independent of light source
	TRI	Terrain roughness Index
	Valley depth	Difference between elevation and ridge
	Visible sky	Percent of visible sky
	VRM	Vector Ruggedness Measure
Spectral bands	Blue	Band 2 (Sentinel)
	Green	Band 3 (Sentinel)
	Red	Band 4 (Sentinel)
	NIR	Band 8 (Sentinel)
	Thermal 1	Band 6 (Landsat)
	Thermal 2	Band 6 (Landsat)
	SWIR	Band 10 (Sentinel)
Spectral indices	BI	$(R\hat{}2 + G\hat{}2 + B\hat{}2)/3\hat{}0.5$
	CI	$(R - G)/(R + G)$
	RI	$R\hat{}2/(B * G\hat{}3)$
	SI	$(R - B)/(R + B)$
Vegetation	Land	Predominant land use
	NDVI	(NIR-R)/(NIR + Red)
	NDWI	(NIR-SWIR)/(NIR + SWIR)
	SAVI	(NIR – R)/(NIR + Red + 0.5)*(1 + 0.5)
Neighbourhood	Latitude	WSG84 coordinates
	Longitude	WSG84 coordinates
Soils	Land types	Original LTS polygons (20 polygons)
	LTS_GM	LTS overlaid with a geomorphon
	Soil forms	South African Soil Classification

An ALOS-2 DEM was obtained from the Japanese Aerospace Exploration Agency (JAXA) to develop 32 covariates representing relief at a 30 m resolution (http://www.eorc.jaxa.jp/ALOS/en). These covariates include local DEM derivatives such as profile curvature, complex DEM derivatives such as the SAGA wetness index (SWI), which were developed in the System for Automated Geoscientific Analysis (SAGA) (Conrad et al., 2015) as well as a geomorphon with a 30 m resolution and a 200 cell search radius. The covariates were thought to represent the relief of the catchment sufficiently.

Spectral images were obtained from both the Sentinel 2A (European Space Agency) from the 11th November 2018 and Landsat 7 from the 16th July 2018 (NASA) satellites (https://earthexplorer.usgs.gov/). Besides the thermal bands of Landsat 7, all spectral bands and indices were developed from the Sentinel 2A Satellite. Covariates which were not at a 30 m resolution were resampled to correspond with the resolution of terrain attributes. This was done using a block averaging resampling method to upscale the Sentinel 2A bands. These bands and indices were thought to represent soil, vegetation, and parent material.

Although the LTS polygons were not used as an input map, the original polygons were used as a covariate and can be seen as representing, soil, macro-climate and relief factors. Additionally, the LTS was overlaid with landform elements which can be seen as representing soils. The neighbourhood factor was represented by latitude and longitude coordinates.

6.3 Results and discussion

6.3.1 Input maps

The TMUs predicted by the five geomorphons are shown in Figure 6.4. These maps were further polygonised to perform the resampled CV test as well as to evaluate the accuracy of predictions. The figure shows that geomorphons with a larger search radii, represent more of a rolling topography by producing larger areas for crest and valley positions (Silva et al., 2016). Additionally, it was observed that geomorphons with a larger radii also produce more continuous polygons which are a better representation of the manually delineated TMUs. This is best seen in valley positions where smaller search radii classify parts of these areas as crest and slope.

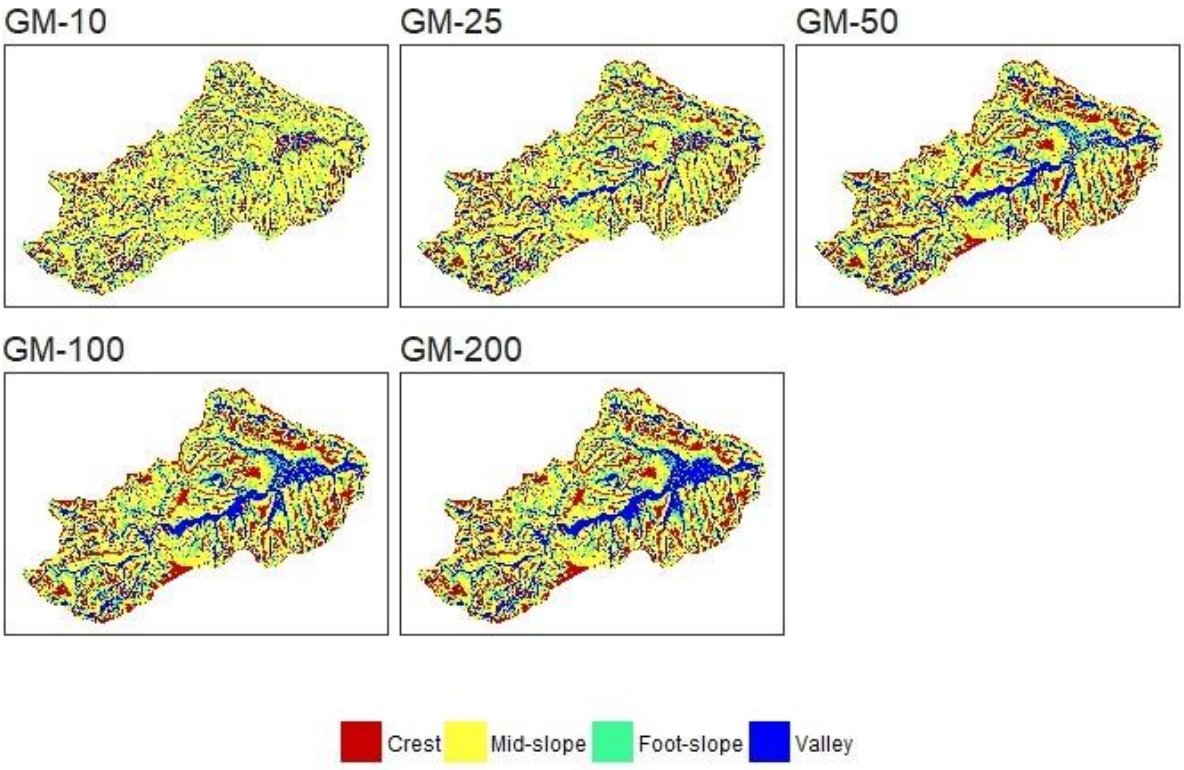


Figure 6.4: Terrain morphological units (TMU) maps predicted through geomorphons with 10, 25, 50, 100, and 200 cell search radii.

The CV tests and predictions through DSMART with 10 realisations for each TMU are shown in Figure 6.5. DSMART was implemented with 50 resamples to predict soil associations and determine if the CV test is a good indication of the best fitting input map. There is a trend from the lowest to the highest search radii according to the resampled CV. A weaker but similar trend can also be seen in the DSMART predictions. In general, the geomorphons with the highest CV, also have the highest prediction potential.

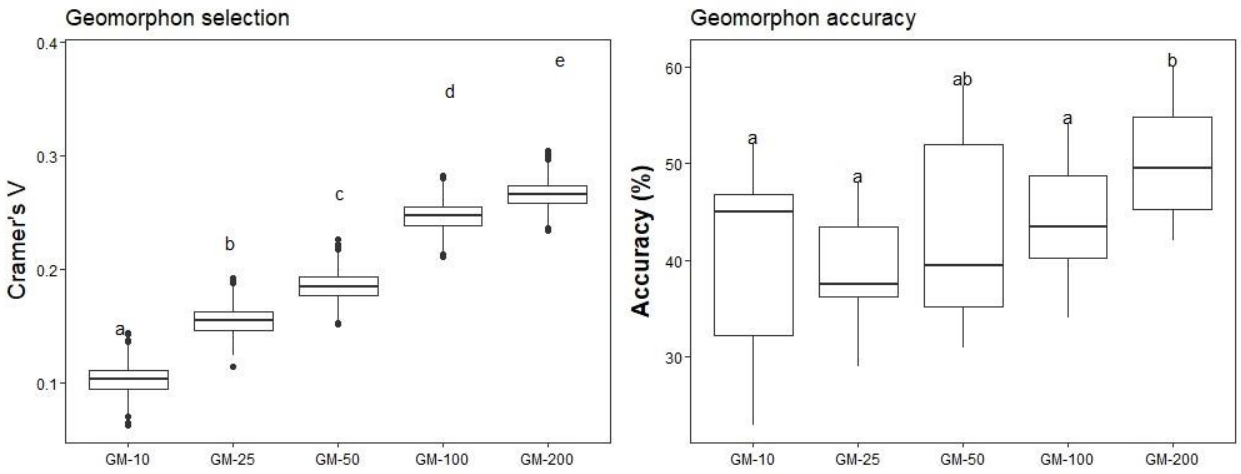


Figure 6.5: Cramer's V and accuracy of soil associations implemented through DSMART for each geomorphon (P < 0.05*).

The GM-200 had the highest CV which was statistically significant relative to all other geomorphons according to a Tukey-Kramer post hoc test. The GM-200 also had the highest prediction potential with a maximum accuracy of 60% when using all 52 covariates. Therefore, the large radius accounts for the variable topography better than the geomorphons with a smaller search radius which create a more planar landscape (flat and mid-slopes). However, the GM-50 produced a statistically similar prediction accuracy to that of the GM-200. These results indicate the CV test can be used as an initial input map selection for DSMART in the Mvoti. Therefore, the GM-200 was selected to test the FSA and the number of resamples. The soil associations and accompanying probability on each TMU for the GM-200 is shown in Table 6.4.

As previously stated, the original LTS polygons were not overlaid with the TMUs as done in Chapters 4 and 5 and when running DSMART, the accuracy decreased when combining the two polygons. For example, when running DSMART on an input map developed by overlaying TMUs (GM-200) on the LTS polygons with 25 resamples per polygon (56 polygons), the average accuracy was 39% with a high of 44%. This result was significantly lower than GM-200 input map with 50 resamples per polygon. This corresponds to the findings by Holmes et al. (2015), who also found that highly detailed input maps can decrease the accuracy of the final predictions. However, this is in contrast to the findings of Møller et al. (2019), who noted that a more detailed map is a better input map.

Table 6.4: TMU units, their percent area, soil associations found on each TMU, and the percent area of soil associations predicted by the LTS.

TMUs	TMU area	Soil association	Soil probability (%)
Crest	25%	Aquic	4
	(78 km^2)	Cambic	6
		Chromic	31
		Leptic	49
		Luvic	10
Mid-slope	38%	Aquic	6
	(120 km^2)	Cambic	13
		Chromic	43
		Leptic	24
		Luvic	13
Foot-slope	12%	Aquic	23
	(37 km^2)	Cambic	31
		Chromic	45
Valley	26%	Aquic	62
	(82 km^2)	Cambic	14
		Chromic	4
		Leptic	7
		Luvic	14

6.3.2 Feature selection and resample size

Model accuracy and REI for all models is shown in Figure 6.6. The highest average accuracy of 50% was achieved when using all covariates with 25 and 50 resamples. The model that uses all covariates with 25 resamples and the RFS with 200 resamples achieved the highest accuracy of 59%. The RFS model had an average accuracy of 49% with 100 resamples while Boruta achieved an average accuracy of 48% (200 resamples) with a high of 57% (50 and 200 resamples). Principle component analysis achieved an average accuracy of 46% (100 and 200 resamples) with a high of 58% (50 and 100 resamples). These results indicate that regardless of the FSA used, the accuracy of predictions will be statistically similar ($p < 0.05$, Tukey-Kramer) to using all covariates. However, and not surprisingly, the control had the lowest average accuracy of 32% with a high of 42% when using 200 resamples per polygon.

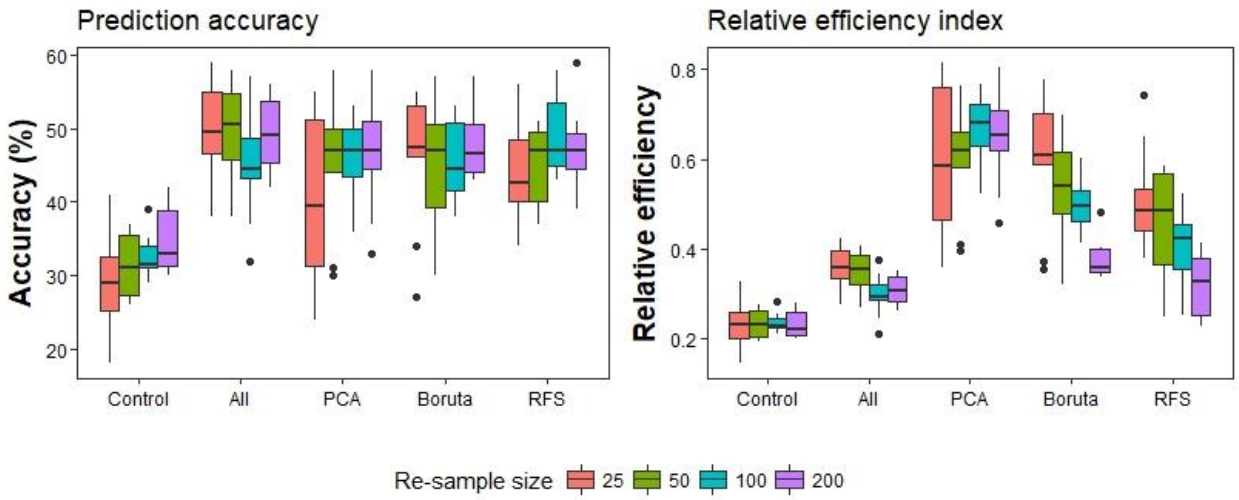


Figure 6.6: Overall accuracy of predictions and relative efficiency index when running DSMART on the control, all 52 covariates (All), Boruta, and RFS selected covariates for 25, 50, 100, and 200 resamples per polygon.

There does not seem to be a trend in resample size and accuracy of predictions. However, there is a trend in the standard deviation and resample size. For example, 25, 50, 100, and 200 resamples (for all models run excluding the control), had average standard deviations of 8.99, 7.37, 5.76, and 5.61, respectively. The larger the number of resamples, the more similar the accuracy will be when running DSMART multiple times. Therefore, larger number of resamples does not necessarily mean greater prediction accuracy but rather, less uncertainty in the accuracy measure. This has implications for running DSMART with only one set of resamples as conducted in this study and has become common with the Random Forest model.

Not surprisingly, PCA (where EI_{max} is 1.63) had the highest average REI (0.63*), followed by Boruta (0.50*), RFS (0.42*), using all covariates (0.32*), and the control (0.23*). The high REI for the PCA models was for two reasons. First, PCA was conducted before running DSMART and therefore, the PCA algorithm was only run once. Second, the first 13 principle components captured 85% of the environmental variability and therefore, only 13 covariates had to be predicted over. The low REI of the control and using all covariates can be accounted for by the large number of covariates which the model had to predict over. Additionally, the control had more polygons to resample. However, this effect was diminished as resample size increased to 200 resamples for Boruta and 100 resamples for RFS.

In addition, there is a clear trend in resample size and REI. This is most clearly seen in the Boruta model where the REI decreases dramatically as resample size increases. This trend is mostly attributed to the larger number of resamples increasing the training set data and therefore, computational time. As resample size increased, so did the number of covariates that were selected by Boruta and RFS which decreased the REI substantially. However, this trend was not seen in the PCA models as the same number of covariates were used during each resample size.

The number of covariates needed to predict over seems to be the biggest limitation to the computational efficiency when running DSMART with FSAs. For example, both supervised FSAs were more computationally efficient than the control for 25 to 100 resamples. However, these models train two different machine learning algorithms compared to the one in the control model. This can be attributed to the ability of RF to generalise the global distribution of soil classes during model training; therefore, most of the computational time is spent on predicting (eager learner) as opposed to training the model (lazy learner) (Liu and Motoda, 1998). This result was unexpected, and the FSAs could be conducted with more iterations to find a better subset of covariates and maintain a competitive runtime.

It is recommended that a compromise be made between prediction accuracy and REI when disaggregating such a complex catchment. However, when disaggregating larger areas, input maps with more polygons or polygon maps with more soil classes, it might be more important to maximize the REI to increase computational efficiency. This is especially true as no model achieved a statistically greater accuracy than another for any resample size. However, the number of resamples should be large enough to have a relatively low standard deviation. Alternatively, weights can be added to the REI for a specific purpose to optimise accuracy, runtime, or the interpretability of the model.

6.3.3 Evaluation of selected model

Both the soil association predictions and confusion index of the selected model are shown in Figure 6.7. Since no model outperformed another in terms of accuracy, the final model was decided on a compromise between accuracy and REI. The final map was produced from the PCA model with 100 resamples. This model achieved an accuracy of 55% with a standard deviation of 4.95%. Additionally, the model achieved an average confusion index of 0.54 and a REI of 0.42 (standard deviation of 0.04) which was the highest efficiency of any model. If a stochastic approach were to be used, the accuracy could increase by 11% when utilising the second most probable class raster.

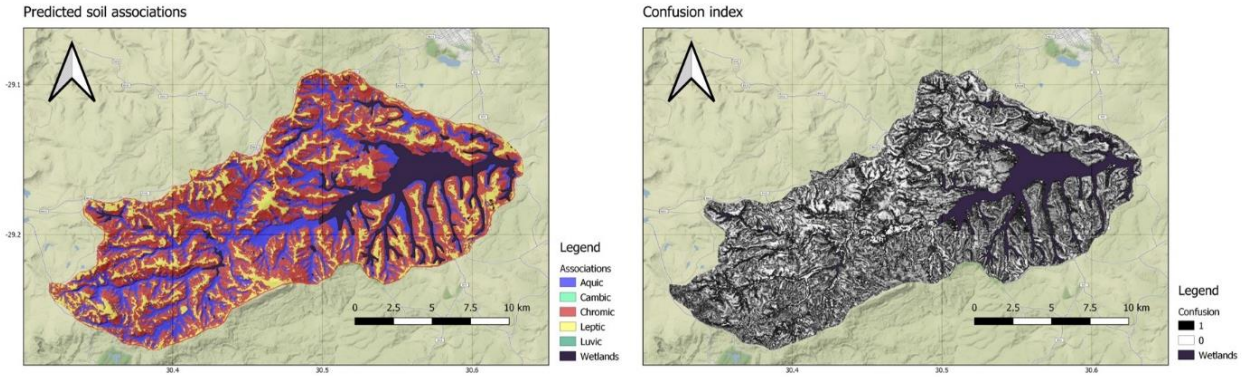


Figure 6.7: Predicted soil associations from the modal of the 10 realisations from the PCA model with 100 resamples per polygon and the confusion between the first and second most probable soil class rasters.

The satisfactory accuracy is due to the model classifying 65% of the chromic association correctly. The chromic association has the highest probability in the LTS legend and was the most observed soil in the external dataset. The model predicted these soils across 56% of the catchment which was not surprising as the Ustic climate and old stable land surface make the Mvoti catchment a conducive environment for ferrallitic processes. However, the model over predicted these soils by 45% according to the LTS.

Low accuracy was achieved for the leptic associations with 37% of the observations correctly classified and were predicted across 19% of the catchment. Therefore, the model underestimated these soils by 42% according to the LTS. The model predicted aquic and cambic soils poorly correctly classifying 10% and 3% of these soils respectively. The poor classification of aquic soils is most likely due to the evaluation sample designs which were predominantly on crest and sloping positions. Additionally, the model did not correctly classify the luvic association most likely because the model only predicted luvic soils at one pixel. The model predicted aquic soils being 24% of the catchment which is similar to the probabilities in the LTS (22%).

The best performing model not predicting cambic and luvic soils well can be attributed to many factors. The low accuracy of cambic and luvic soils could be attributed to these soils being classified as aquic soils. Additionally, cambic and luvic soils were also classified as chromic soils. Therefore, it was thought that by splitting the chromic association into rhodic and xanthic associations, it might increase the model's ability to predict soils with low probabilities as well as produce a more detailed

soil map. However, this was not the case when running DSMART with all covariates and 100 resamples. Instead, the average accuracy dropped to 18% and cambic and luvic soils were still not predicted well. This indicates that increasing the complexity dramatically decreases the accuracy of DSMART and does not improve the predictions of soil associations with low probabilities.

The poor predictions of cambic or luvic soils could be attributed to the ensemble method of Random Forest. The majority vote of Random Forest predictions makes it more difficult for the algorithm to predict classes with a low probability. This could be accounted for by using a different algorithm. For example, when using linear discriminatory analysis with 50 resamples, which predicted all of the classes and had an average accuracy of 45% which was statistically the same as Boruta and RFS model with the same resample size. This warrants the use of other machine learning algorithms in DSMART as conducted in Chapter 5.

Alternatively, the low accuracy of these associations could be due to inaccuracies in the LTS legend, or the imbalanced datasets in both the evaluation observations and LTS legend as well as the sample design. This could potentially be corrected for using upsampling as done in Chapter 5, however, this would have inhibited the evaluation of resample size as the size as the soil allocation will be different for each realisation.

Surprisingly, the model had a rather modest confusion index of 0.54. The lowest confusion was seen in valley positions were aquic soils are prevalent and have a high probability in the LTS legend. Low confusion was also seen on crest positions where leptic soils have a high probability of occurrence. The highest confusion was seen on low lying sloping positions where all soils had a modest probability. This is not surprising as cambic and luvic soils were classified with a low accuracy due to misclassification of aquic and chromic soils.

6.3.4 Comparison with Land Type Survey

Soil associations overlaid with the original LTS polygons is shown in Figure 6.8. As expected, it is clear the LTS surveyors used TMUs to estimate the soil distribution. For example, there are clear toposequences in each land type. In general, leptic soils are on crests, chromic soils on mid-slopes, periodic cambic and luvic soils on foot-slopes, and aquic soils in valley positions. This trend can be seen throughout the catchment but is most clearly expressed on larger land types while some smaller land types do not show the full catena sequence. Some land types that are in lower elevation

positions at the bottom of the catchment (Eastern part) do not show this trend because these positions do not have all TMUs. For example, land type Bb106 only has foot-slopes and valleys. Therefore, this land type consists of aquic soils and wetlands.

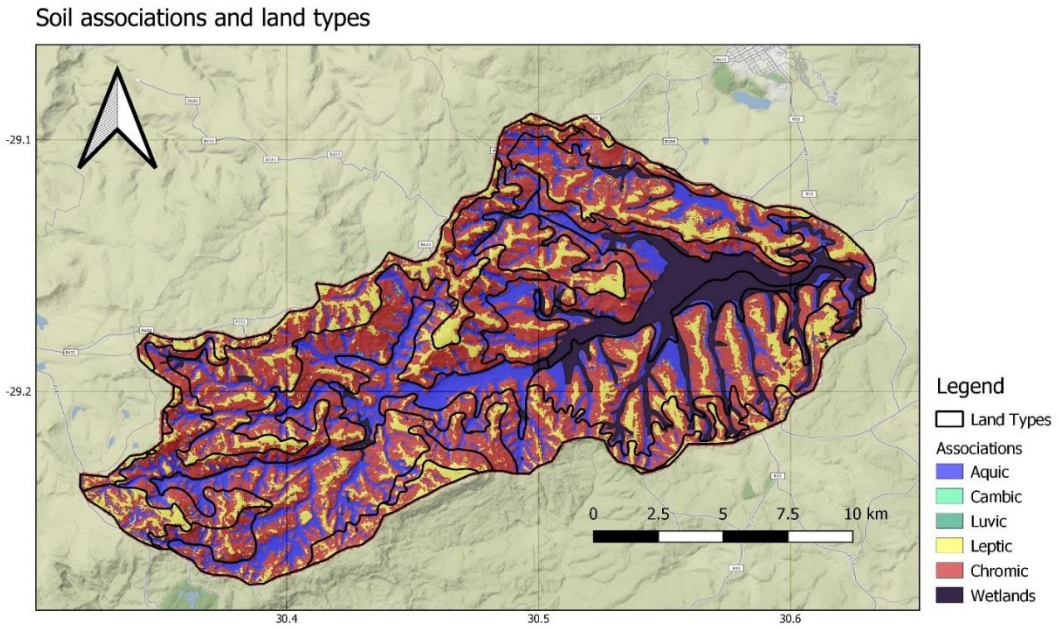


Figure 6.8: Soil associations overlaid with the original Land Type Survey polygons.

6.3.5 Covariate importance

The covariate importance of the final model is shown in Table 6.5. The covariate importance is the mean decrease in accuracy when a particular covariate is removed from the model averaged over the 10 realisations. This is calculated from the out-of-bag error of each Random Forest. PC2 had the highest overall importance but also the highest standard deviation. It seems the more important the covariate is, the higher standard deviation it will have. Therefore, PC2 importance varied widely for each tree grown. It was thought that PC1 would be the most important covariate as it accounted for the most variation (22%) of the data while PC2 accounted for 16% of the data.

The variables which correlate most to PC2 were slope curvature (local and regional), relative elevation, landform elements, mass balance index and the SWI. Variables which correlate most with PC8 were analytical hillshading, aspect, flow direction, gradient difference, mid-slope position, and soil forms. Therefore, the soil associations orthogonally correlate to a variety of covariates mostly related to the relief factor but also the soil factor.

Table 6.5: Covariate importance showing mean decrease accuracy, standard deviation (Sd), min and max values, and range.

Rank	PCA	Mean	Sd	Min	Max	Range
1	PC2	55	32.7	1.52	100	98
2	PC8	26	10.7	0.00	57	57
3	PC9	23	12.5	3.41	61	58
4	PC10	22	13.1	3.15	68	64
5	PC1	22	10.9	0.00	49	49
6	PC3	21	10.7	0.58	47	46
7	PC5	21	11.6	0.00	60	60
8	PC12	21	9.9	2.70	48	45
9	PC11	20	10.7	0.00	48	48
10	PC6	20	8.5	4.83	39	34
11	PC7	20	11.3	0.00	52	52
12	PC4	19	10.3	0.00	47	47
13	PC13	18	8.6	3.91	43	39

Latitude and land types correlated to PC10 while spectral bands (BI, SI, red) as well as vegetation (SAVI, NDVI) correlate to PC1. Although these covariates influenced the predictions, it was surprising that they did not have a greater influence. For example, the soils varied greatly from the top to the bottom (west to east) of the catchment (neighbourhood). Additionally, forest plantations are generally at the top of the catchment on crest to sloping positions (vegetation). Therefore, it was thought that neighbourhood and vegetation would have a larger influence on the soil distribution.

6.4 Conclusion

This study aimed to introduce a method to quantitatively select input maps and introduce three feature selection approaches when running DSMART to disaggregate complex soil-terrain polygons. The input maps were selected through a spatially resampled CV test. Feature selection algorithms were compared by their overall accuracy, runtime, as well as their ability to reduce data dimensionality evaluated through a REI. Additionally, a different number of resample sizes were compared for each FSA to determine the optimal number of resamples. The FSAs were compared against using 52 covariates when disaggregating the original LTS polygons.

It was found that the resampled CV test is a good indication of the optimal input map. This was verified by the prediction accuracy of each input map. It was shown that geomorphons with larger search radii, in general, have a larger CV and produce greater accuracy through DSMART. When

testing DSMART with all covariates and the three FSAs, no statistical significance was found for prediction accuracy. However, there was a significant drop in accuracy when only disaggregating the original LTS polygons. Additionally, the FSAs had a higher REI than when using all covariates and when running the control. This indicates that FSAs are more computationally efficient and use less covariates, therefore FSAs are more interpretable.

The findings from this study are applicable when disaggregating larger areas or other datasets as it introduces methods which maintain an compared to using a large number of covariates but increase computational efficiency. Further work should include approaches to handle imbalanced datasets both in the soil legacy legends and field observations. This could improve the accuracy of DSMART especially for soil classes with a low probability of occurrence. Additionally, further work should focus on increasing the detail at which DSMART can predict soil classes to produce functional soil maps.

6.5 References

Breiman, L., 2001. Random Forests. Berkeley, California. https://doi.org/10.1017/CBO9781107415324.004

Breiman, L., 1996. Bagging predictors. Mach. Learn. 2, 123–140.

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

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analysis (SAGA). Geoscientific Model Development. https://doi.org/doi:10.5194/gmd-8-1991-2015

GRASS Development Team, 2017. Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Found.

Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues. Soil Sci. Soc. Am. J. 75, 1201. https://doi.org/10.2136/sssaj2011.0025

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. Springer Series in Statistics.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., De Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One 10, 1–26. https://doi.org/10.1371/journal.pone.0125814

Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area spatial disaggregation of a mosaic of conventional

soil maps: evaluation over Western Australia. CSIRO 53, 865–880.

Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons - a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156. https://doi.org/10.1016/j.geomorph.2012.11.005

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97, 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, the R.C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., 2018. Caret: Classification and Regression Training.

Kursa, M.B., Rudnicki, W.R., 2010. Feature Selection with the Boruta Package. J. Stat. Softw. 36, 293–327. https://doi.org/Vol. 36, Issue 11, Sep 2010

Land Type Survey Staff, 1972-006. Land Types of South Africa on 1:250 000 scale. Pretoria, South Africa.

Larose, D.T., 2006. Dimension Reduction Methods, in: Data Mining Methods and Models. John Wiley & Sons, Inc, pp. 1–32. https://doi.org/10.1002/0471756482.ch1

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Springer Science+Business Media, LLC, New York.

McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. Nutr. Cycl. Agroecosystems 50, 51–62. https://doi.org/10.1023/A:1009778500412

McBratney, A.B., Santos, M.L.M., Minasnya, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

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

Møller, A.B., Malone, B., Odgers, N.P., Beucher, A., Vangsø, B., Humlekrog, M., Minasny, B., 2019. Improved disaggregation of conventional soil maps. Geoderma 341, 148–160. https://doi.org/10.1016/j.geoderma.2019.01.038

Nel, J.L., Driver, A., Strydom, W.F., Maherry, A., Petersen, C., Hill, L., Roux, D.J., Nienabar, S., van Deventer, H., Swartz, E., Smith-Adao, L.B., 2011. Atlas of freshwater ecosystem priority areas in South Africa: Maps to support sustainable development of water resources. Pretoria, South Africa.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100. https://doi.org/10.1016/j.geoderma.2013.09.024

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc, San Francisco, California.

R Core Team, 2017. R: A language and environment for statistical computing.

Ranst, E. Van, Verdoodt, A., Baert, G., 2010. Soil Mapping in Africa at the Crossroads: Work to Make up for Lost Ground, in: Section of Natural and Medical Sciences. Ghent University, Ghent, pp. 147–163.

Silva, S.H.G., Menezes, M.D., Mello, C.R., Góes, H.T.P., Owens, P.R., Curi, N., 2016. Geomorphometric tool associated with soil types and properties spatial variability at watersheds under tropical conditions. Sci. Agric. 73, 363–370. https://doi.org/10.1590/0103-9016-2015-0293

Soil Classification Working Group, 1991. Soil Classificatioin: a Taxonomic System for South Africa, 2nd ed. Department of Agricultural Development, Pretoria, South Africa.

Soil Survey Staff, 2014. Keys to soil taxonomy, 12th ed. USDA-Natural Resources Conservation Service, Washington, DC.

van Zijl, G., 2019. Digital soil mapping approaches to address real world problems in southern Africa. Geoderma 337, 1301–1308. https://doi.org/10.1016/j.geoderma.2018.07.052

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2016. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142. https://doi.org/10.1016/j.geoderma.2016.06.006

Wiese, L., Ros, I., Rozanov, A., Boshoff, A., de Clercq, W., Seifert, T., 2016. An approach to soil carbon accounting and mapping using vertical distribution functions for known soil types. Geoderma 263, 264–273. https://doi.org/10.1016/j.geoderma.2015.07.012

Zeraatpisheh, M., Ayoubi, S., Brungard, C., Finke, P., 2019. Disaggregating and updating a legacy soil map using DSMART, fuzzy c-means and k-means clustering algorithms in Central Iran. Geoderma 340, 249–258. https://doi.org/10.1016/j.geoderma.2019.01.005

Chapter 7 Recommendations and further work

7.1 Conclusion

Digital soil mapping offers a quantitative method to predict soil classes and soil properties. This allows soil maps to be reproduceable and can display continuous soil properties. Soil information is important for land use management for both smallholder and commercial farms as well as environmental management as accurate soil information is required for such things as hydrological models. However, due to limited resources, highly detailed spatial soil information is sparse in South Africa. Therefore, increasing the accessibility and accuracy of spatial soil information is worthy of further investigation.

This research aimed to develop two DSM frameworks using resources available in South Africa and freely available technologies/software. Starting in the Swartland, Western Cape, Framework 1 aimed to produce farm-scale soil maps and patterns using point observations consisting of two objectives. Objective 1 was to produce farm-scale soil maps of multiple soil properties by simultaneously optimising FSAs and predictive models. Objective 2 was to derive soil patterns through an aggregated geomorphon as an initial indication of spatial soil variability.

The framework achieved the objectives in many ways. Objective 1 produced farm-scale soil maps of five different soil properties and achieved results comparable to other studies. It was shown that Boots feature selection together robust linear regression proved to be the most accurate model for 4 out of 5 soil properties. Objective 2 quantitatively selected a geomorphon which stratified the soil-landscape for multiple soil properties. The aggregation of geomorphon units was based on the distribution of soil associations through a decision tree. This produced meaningful soil patterns and gave pedological insight into the soil distribution on the site. These two objectives represent a step towards increasing soil information accessibility and accuracy to farmers.

Framework 2 developed a methodology to disaggregate the LTS from the farm-scale in the Sandspruit to the land type scale in KwaZulu Natal and Eastern Cape, and finally to the catchment-scale in KwaZulu Natal, Midlands. Framework 2 consisted of three objectives. Objective 3 was to disaggregate the LTS into a farm-scale soil depth class map utilising the DSMART algorithm. Objective 4 compared 10 algorithms implemented through DSMART in two environmentally contrasting locations (Cathedral Peak and Ntabelanga) at the land type scale. Objective 5 was to

further optimise DSMART by providing a rapid input map selection method, incorporating FSAs into DSMART, and by analysing resample size at the catchment scale (Mvoti catchment).

Objective 3 downscaled the small-scale LTS polygons into a soil depth class map through a two-step disaggregation approach. This produced a soil depth class map with a 68% accuracy through MLR. As far as this study is aware of, this is the first study which downscaled a national resource map to the farm-scale as well as predicted soil depth classes through DSMART. Objective 4 increased the usability of DSMART by allowing additional algorithms to be run. It was found that regularised linear models worked best at the land type scale. This allows DSMART to be more versatile as the user can select and optimise the predictive model. Objective 5 simultaneously disaggregated 20 land types into five soil associations and developed additional features which maximise the computational efficiency of DSMART. The final map produced an accuracy of 55% through PCA and RF predictions. This is a major advancement, as past studies have only disaggregated one or two land types. These three objectives contribute to aiding in both farm and environmental soil acquisition. Another major benefit of this approach is that it eliminates much of the subjectivity when disaggregating the LTS. Furthermore, the semi-automated framework produces a more rapid method of disaggregation than an expert knowledge approach.

7.2 Limitations and recommendations

Framework 1 developed in Objective 1 was determined to be a robust approach. However, the fact that the model residuals showed spatial autocorrelation and kriging improved the model, indicates that the FSA-predictive models were still biased and can be misleading. Therefore, it is recommended that the hybrid approach be used which requires knowledge of both geostatistics, machine learning, and soil science. Objective 2 showed that quantitative selection and aggregation of geomorphons can stratify many soil properties. However, the GM-5 did not stratify all soil properties (e.g., EC) and the manually delineated LFE system was a better predictor for EC, SOC, and ERD. Therefore, it is difficult to determine which LFE classification system depending on soil property of interest. It is recommended that multiple LFE classification systems be used to determine the appropriate system for each soil property.

Framework 2 proved to be able to extract the expert knowledge used in developing the LTS to produce more detailed soil maps on multiple scales. However, there are three large limitations that

were observed. Imbalanced data in either the LTS or evaluation data have a large effect on model accuracy. It is recommended that soil classes be grouped into soil associations to account for these imbalances as well as for differences in the LTS and evaluation data (soils classified differently). In regions with low relief and complex geology, the method produced unsatisfactory results. Therefore, when available, a geological map should be incorporated, and expert rules should be implemented into DSMART. Although methods were introduced to increase computational efficiency, the computational power needed is large. The combination from data input, data preparation, and running the DSMART model all contribute to computational inefficiency of the method. Furthermore, the framework is still relatively inaccessible due to the LTS not being freely available and the relatively lack of pedometric knowledge in South Africa.

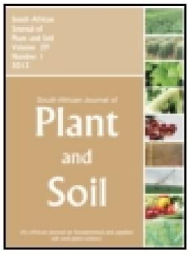
7.3 Further work

Although a step towards solving agriculture and environmental management issues in South Africa. There is further work that is required to improve land use optimisation and environmental models. For example, mapping of soil attributes at different depths is required to get a better understanding of soil subsurface properties that influence such things as ERD and water redistribution. This is applicable to both Framework 1 and 2.

The tools are available to disaggregate the LTS on the provincial and even country scale. However, this would involve four additional contributions. First, access to the LTS needs to be improved because as of now, it is not freely available outside of academia. Therefore, industry has very little access to the LTS. Second, evaluation points would have to be obtained, most likely from many different data sources. Therefore, collaboration between government agencies and industry is needed to produce and evaluate the models. Third, the LTS legend would have to be changed into a format suitable to be run through DSMART. This would involve a large team and many man hours to convert the data for the whole country. Forth, a detailed geological map would be needed especially to map the interior of South Africa in areas where the soils are controlled by parent material.

At present, it is difficult to determine if DSM can address the needs for a large population in South Africa. However, it is a tool which can be advanced to address these issues. Therefore, DSM should be further pursued. In doing so, it will also increase the knowledge of the spatial soil distribution throughout South Africa which can be applied to address many issues such as economic, food, and water security.

Appendix A



South African Journal of Plant and Soil

ISSN: 0257-1862 (Print) 2167-034X (Online) Journal homepage: https://www.tandfonline.com/loi/tjps20

High-resolution digital soil mapping of multiple soil properties: an alternative to the traditional field survey?

Trevan Flynn, Willem de Clercq, Andrei Rozanov & Cathy Clarke

To cite this article: Trevan Flynn, Willem de Clercq, Andrei Rozanov & Cathy Clarke (2019): High-resolution digital soil mapping of multiple soil properties: an alternative to the traditional field survey?, South African Journal of Plant and Soil, DOI: 10.1080/02571862.2019.1570566

To link to this article: https://doi.org/10.1080/02571862.2019.1570566

South African Journal of Plant and Soil 2019: 1–11
Printed in South Africa — All rights reserved

© Southern African Plant & Soil Sciences Committee
SOUTH AFRICAN JOURNAL OF PLANT AND SOIL
ISSN 0257-1862 EISSN 2167-034X
https://doi.org/10.1080/02571862.2019.1570566

High-resolution digital soil mapping of multiple soil properties: an alternative to the traditional field survey?

Trevan Flynn[1], Willem de Clercq[2], Andrei Rozanov[1] and Cathy Clarke[1]*

[1] *Department of Soil Science, Stellenbosch University, Stellenbosch, South Africa*
[2] *Stellenbosch Water Institute, Stellenbosch University, Stellenbosch, South Africa*
* *Corresponding author, email: cdowding@sun.ac.za*

Spatial information on soil particle size distribution and soil organic carbon (SOC) are important for land-use management, environmental models and policy-making. Digital soil mapping (DSM) techniques can quantitatively predict these soil properties using minimal resources. However, DSM has not been adequately evaluated at the farm-scale. The aim of this study was to optimise the DSM framework to produce farm-scale soil maps for 366 ha in the Sandspruit catchment, Western Cape, South Africa. Four feature selection techniques and eight predictive models were evaluated on their ability to predict particle size distribution and SOC. A boosted linear feature selection produced the highest accuracy for all but one soil property. The top-performing predictive models were robust linear models for gravel (ridge regression, RMSE 9.01%, R^2 0.75), sand (support vector machine, RMSE 4.69%, R^2 0.67), clay (quantile regression, RMSE 2.38%, R^2 0.52) and SOC (ridge regression, RMSE 0.19%, R^2 0.41). Random forest was the best predictive model for silt content (RMSE 4.12%, R^2 0.53). This approach appears to be robust for farm-scale soil mapping where the number of observations is often small but high-resolution soil data are required.

Keywords: digital soil mapping, farm-scale, feature selection, high resolution, machine learning

Online supplementary material: Supplementary information for this article is available at https://dx.doi.org/10.1080/02571862.2019.1570566

Introduction

Spatial information on soil particle size distribution (PSD) and soil organic carbon (SOC) is increasingly important for land-use management, environmental models and policy-making. However, to obtain such information is time consuming, costly, and the existing (mostly small-scale) soil maps lack the necessary detail and resolution. Digital soil mapping (DSM) offers a way to quantitatively predict soil properties in a cost-effective manner. However, DSM techniques are seldom used at the farm-scale due to a lack of expertise and financial resources in the industry (Xu et al. 2018).

Developed by McBratney et al. (2003), the *scorpan* method is the most widely used DSM framework and has facilitated the growth of DSM applications in recent years (Minasny and McBratney 2016). The *scorpan* method uses machine learning models to find soil environmental relationships and predict soils in unobserved locations. The method predicts soil attributes as a function of the five soil-forming factors, additional soil properties, and location. The inclusion of additional soil properties and location indicates that soils can be predicted through other (easily measured) soil properties (McBratney and Webster 1983), geographic location (Lagacherie 1992), and location relative to important features such as distance to stream (Bui and Moran 2000). These *scorpan* factors are obtained from remote sensing, proximal sensing or easily measured soil properties known as covariates. This approach has been added to the US Department of Agriculture soil survey manual as it becomes more available and necessary (Soil Survey Staff 2017).

There are many predictive models, such as various types of linear regression (McKenzie and Austin 1993; Chagas et al. 2016), decision trees (Moran and Bui 2002; Jafari et al. 2014; Subburayalu et al. 2014), random forest (Pahlavan et al. 2014; Hengl et al. 2015; Chagas et al. 2016), generalised additive models (Bishop and McBratney 2001), artificial neural networks (Behrens and Förster 2005; Brungard et al. 2015; Aitkenhead and Coull 2016), and fuzzy logic models (De Gruijter et al. 1997; Lagacherie et al. 1997; Zhu 1997; Qi et al. 2006; van Zijl et al. 2013). There has been much research into predictive models, however, many of the above-mentioned studies conclude that there is no single algorithm that predicts all soil properties best. Therefore, many predictive models must be tried to obtain the best accuracy for the soil property of interest (Kuhn and Johnson 2013).

Although there are many predictive models and a common DSM framework, the *scorpan* method does not explicitly place feature selection into the framework as the method focuses on capturing all the *scorpan* factors. Due to this, feature selection is not a commonly researched aspect in DSM (Behrens et al. 2010) and McBratney et al. (2003) state that there is a need for more research in this regard. Furthermore, many authors note that different

and/or additional covariates could have improved the model and, at the farm-scale, not all *scorpan* factors are needed due to low variance in their values (e.g. climate) over short distances. Commonly used feature selection techniques include principle component analysis (PCA; Hoffmann et al. 2014), ANOVA analysis (Sun et al. 2011), step-wise linear reduction (Mora-Vallejo et al. 2008) and recursive feature selection (Brungard et al. 2015). It is hypothesised that optimising and treating feature selection techniques and predictive models simultaneously is a robust approach suitable for farm-scale soil mapping.

The aim of this paper was to evaluate a DSM framework to predict soil properties at the farm-scale in the Sandspruit catchment, South Africa. This DSM framework can be seen as an adaptation of the *scorpan* method and specifies the evaluation of feature selection and predictive models simultaneously. This is important where bias and collinearity extensively affect small data sets such as farm-scale soil surveys (Kuhn and Johnson 2013). This can affect the interpretability of the model as the model coefficients and the standard error may be misleading (Mason and Perreault 1991).

Materials and methods

Site location

The research site lies at approximately 33°14′51.72″S, 18°08′52.52″ E in the middle of the Sandspruit river catchment, Western Cape, South Africa. The catchment location and sample design are shown in Figure 1a and 1b, respectively. The site was chosen to capture as much landscape variability as possible at the farm-scale. The area covers 366 ha under dry-land agriculture with wheat and canola crop rotation. The altitude ranges between 94 m and 220 m above sea level. The predominant geology of the landscape is greywacke, phyllite, and schist of the Moorreesburg, Klipplaat, and Berg River formations of the Malmesbury group. There are silcrete and ferricrete outcrops near the site but not directly on the area of interest. These outcroppings separate the old African surface and the younger dissected land surface at lower altitudes.

In total, 93 soil profiles were classified and sampled, which were used for both training and validating the models. There were two sampling schemes developed to sufficiently cover terrain. The two sampling schemes were expert samples and conditioned Latin hypercube sampling (cLHS). The expert (Dr Freddie Ellis, with 50 years mapping experience in the region) used 5 m contour lines overlaid on a Google Earth image to place profiles that would capture maximum soil variation.

Conditioned Latin hypercube sampling is a type of random sampling strategy that stratifies the samples on the values of multiple covariates (Minasny and McBratney 2006). This captures the environmental variation and terrain dissection. The cLHS was implemented with 100 000 iterations on six environmental covariates. The covariates were chosen based on previous studies and included aspect, compound topographic index, plan curvature, profile curvature, slope and soil adjusted vegetative index.

Soil properties

Five topsoil properties were measured (composition samples) to assess the DSM framework. The descriptive statistics and frequency distribution of the soil properties are shown in Table 1 and Figure 2, respectively. All soil properties were measured after air drying and sieving the soil (<2.0 mm). Gravel content was measured by taking the gravimetric weight percent of the coarse fragments. Sand, silt and clay contents were measured by the pipette method and the sand grade was determined by the sieve method (Soil Survey Staff 2014). The SOC was measured using the Walkley–Black method (Walkley and Black 1934) to determine the SOC percentage.

Digital soil mapping framework

A diagram of the adapted DSM framework is shown in

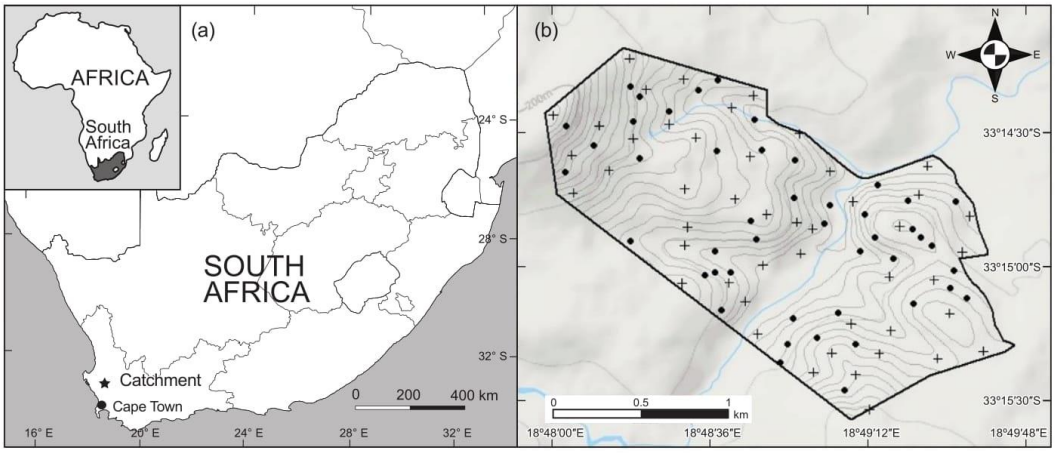


Figure 1: The catchment location within South Africa (a) and the site showing the soil sample design (b)

Figure 3. A pool of covariates was developed from which feature selection techniques selected covariates. Each combination of covariates, selected by each feature selection technique, was used to predict each soil property. This was an iterative process for all feature selection and predictive model combinations. In total, 320 models were run, and 64 models were run for each soil property. The best-performing model was determined by the highest accuracy of predictions (root mean square error; RMSE). Due to the limited number of samples, the models were validated through leave-one-out cross-validation. Leave-one-out cross-validation leaves one observation out, trains the model and predicts on the single left-out sample. This is done for all observations and the accuracy is averaged.

The mathematical representation of this DSM framework is shown in Equation 1:

$$S = f(g(Q)) + \in \quad (1)$$

where S is the soil property of interest, the function f is a deterministic predictive model that establishes soil environmental relationships with the covariates Q, selected from the feature selection function g, and $\in$ is the independent random error added to the model. The equation assumes that the covariates Q are developed from all covariates that can be obtained regardless of the expert opinion. In other words, the expert does not know the true relationship between soil properties and the covariates. The $\in$ term is stochastic and is defined by a sample variogram. The sample variogram is used to krige the residuals and the values are added to the trend calculated by function f.

The difference between this method and the *scorpan* method is that quantitative feature selection is specified and optimised together with the predictive model. This approach increases the importance of which covariates are placed into the predictive model as the function f, which cannot be defined until after the function g, selects appropriate covariates. This specification decreases the subjectivity and collinearity of covariate selection, which is important for small data sets such as a farm-scale soil surveys (Mason and Perreault 1991). In addition, this can add to the interpretability of the model as more emphasis is placed on the covariates.

Table 1: Soil property descriptive statistics showing the mean, SD and range

Soil property	Mean	SD	Range
Gravel (%)	39.9	18.3	0.0–64.0
Sand (%)	60.3	9.2	19.6–76.7
Silt (%)	29.7	7.0	16.5–65.6
Clay (%)	10.1	3.7	3.88–22.7
Soil organic carbon (%)	0.65	0.7	0.07–1.20

Primary data sources

A 30 m digital elevation model (DEM) was acquired from the Advanced Land Observation Satellite (ALOS-2) provided by the Japanese Aerospace Exploration Agency (JAXA; http://www.eorc.jaxa.jp/ALOS/en/aw3d30). The DEM was obtained from the S034E018 thumbnail corresponding to the geographic coordinates of 34° S, 18° E. The DEM was selected because the vertical

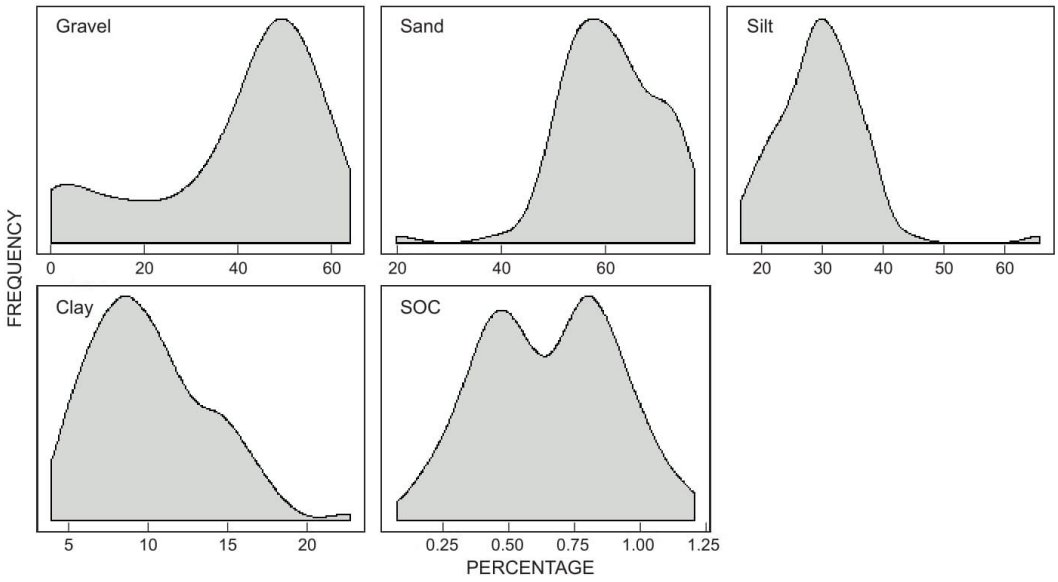


Figure 2: Soil property distribution based on 93 soil observations. SOC = soil organic carbon

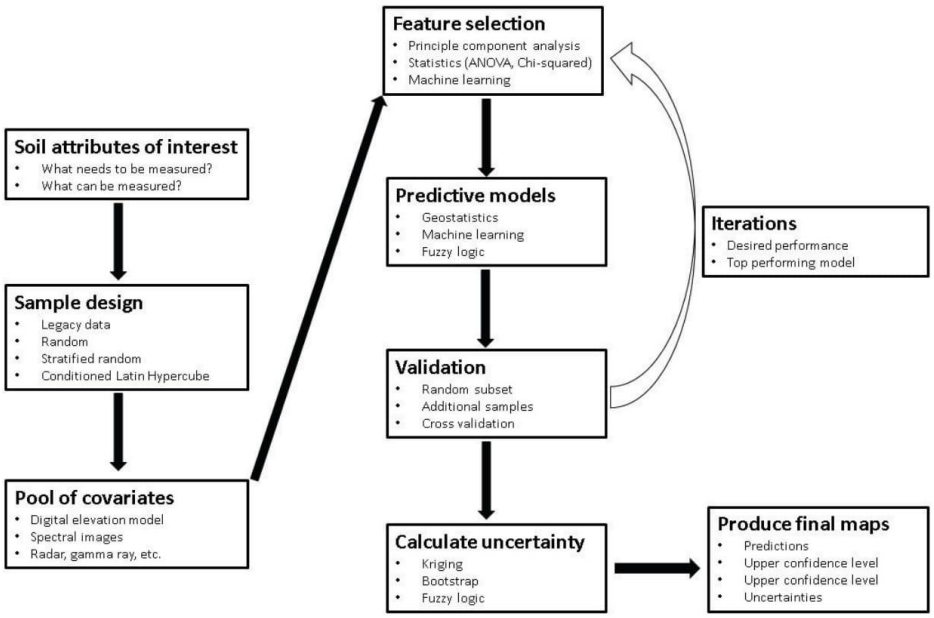


Figure 3: Flow diagram of the digital soil mapping methodology

accuracy (6 m) is significantly higher than other freely available DEMs (e.g. 30 m SRTM) at the same resolution.

Sentinel-2A images with four spectral bands were used for extraction of indices described below.

Covariate development

A pool of covariates was developed to capture organisms, relief, parent material and age according to the *scorpan* factors. The covariates were easily calculated topographic derivatives, and spectral images that were thought to capture the environmental variation sufficiently. In total, 47 topographic covariates and 38 spectral covariates were developed. Many covariates (relative to soil observations) were used to decrease subjectivity in covariate selection and evaluate feature selection techniques. The resolution of the covariates was defined by the finest legible resolution according to inspection density for 366 ha and 93 soil observations (Hengl 2006). The equation for the finest legible resolution is shown in Equation 2:

$$\text{Finest resolution} = 0.05 \times \sqrt{\frac{A}{N}} \quad (2)$$

where A is the square meters of the area and N is the number of soil observations.

Topographic covariates are shown in Table 2. A description of all topographic covariates can be found in Hengl and Reuter (2009). Covariates that might be redundant (e.g. standardised height and normalised height) were not removed so as to evaluate the feature selection techniques. All spectral covariates, a description of the covariates and their associated calculations are listed in Table 3. A description of the spectral covariates can be found in Huete (1988), and Ray et al. (2004). The spectral bands and indices were selected to incorporate soil, age, parent material and vegetation according to the *scorpan* method.

Feature selection

In total, five feature selection techniques were evaluated as shown in Table 4. Feature selection techniques were selected to incorporate a variety of algorithms that have little or no tuning parameters. This includes two linear models and two random forest models. A short description of how each algorithm was implemented can be found in Supplementary Appendix S1, and a detailed description of each feature selection technique can be found in Hastie et al. (2009) and Kuhn and Johnson (2013).

Predictive models

The predictive models implemented are shown in Table 5. A general description of each model can be found in Kuhn and Johnson (2013). The models were chosen to get a wide range of machine learning techniques from simple to complex. Simple linear models include ridge regression (RR), linear boosted models (LBM) and quantile regression (QR). Non-linear models include support vector machines (SVM), random forest (RF), stochastic gradient boosting (SGB), cubist (CB) and penalised additive spines (P-splines). For a detailed description of the implementation of each algorithm see Supplementary Appendix S1.

Table 2: All topographic covariates derived from the ALOS-2 digital elevation model and their SAGA module representation

Representation	Covariate	Representation	Covariate
Land form elements	Land type geomorphon (LTS-GM)	Lighting/exposure	Analytical Hill shading
Hydrology characteristics	Channel network base level		Diffuse insolation
	Flow directions		Direct insolation
	Catchment area		Negative openness
	Catchment slope		Positive openness
	Flow path length		Sky view factor
	Modified catchment area		Visible sky
	Slope length	Local morphometry	Flow line curvature
	Slope length factor (LS factor)		Plan curvature
	Stream power index		Profile curvature
	SAGA wetness index (SWI)		Tangential curvature
	Topographic wetness index		Total curvature
	Melton ruggedness index		Aspect (degrees)
Landscape morphometry	Maximum height		Convergence Index
	Mid-slope position		Convexity
	Multiresolution ridge top flatness index (MRRTF)		Cross-section curvature
			Elevation
	Multiresolution valley bottom flatness index (MRVBF)		General curvature
			Downslope gradient
	Normalised height		Longitudinal curvature
	Slope height		Mass balance index
	Standardised height		Maximum curvature
	Topographic position index		Slope (degrees)
	Valley depth		Terrain ruggedness index
			Vector ruggedness index (VRI)

Table 3: Soil and vegetative bands and indices used for soil spatial variability analysis

Band	Band length	Symbol	Resolution
Red	0.665	Red	10 m
Near infrared (NIR)	0.842	NIR	10 m
Short-wave infrared 1	1.610	SSWIR	20 m to 10 m
Short-wave infrared 2	2.190	LSWIR	20 m to 10 m
Index	**Calculation**	**Property**	**Resolution**
Brightness index	$(R^2 + G^2 + B^2)\ /3^{0.5}$	Average reflectance	10 m
Colouration index	$(R - G)/(R + G)$	Soil colour	10 m
Redness index (RI)	$R^2/(B \times G^3)$	Hematite	10 m
Saturation index (SI)	$(R - B)/(R + B)$	Spectral slope	10 m
Soil adjusted vegetative index	$\frac{NIR - R}{NIR + R + 0.1}(1 + 0.1)$	Chlorophyll reflectance	10 m

Spatial autocorrelation

Spatial autocorrelation was evaluated through sample variogram analysis. Due to the small sample size, only isotropic variograms were analysed as there were not enough point-pairs for an anisotropic variogram (Webster and Oliver 2001). The variograms were estimated through residual maximum likelihood analysis (REML) as this is considered best practice (Lark et al. 2006). REML performs general least squares regression but corrects the residuals by maximising the likelihood measure. Residuals of each model were evaluated for normality before variogram analysis and the kriged residuals were added back to the trend of the deterministic model. The benefit of REML is that the variance term does not need to be stationary and the covariates can be correlated in space and time. In addition, a variogram fitted through REML needs less samples than other least squares methods (Kerry and Oliver 2007).

Table 4: Feature selection algorithms used for both regression and classification models

Technique	Type	Algorithm
Univariate	Filter	Random forest
Recursive	Wrapper	Random forest
LASSO	Embedded	Regularised linear
Boost	Embedded	Regularised linear

Spatial uncertainties

Prediction ranges were estimated by developing fuzzy *k*-means clusters with extragrades (FKMe) of covariates

with similar distribution of model errors (McBratney and De Gruijter 1992). The FKMe algorithm is similar to fuzzy *k*-means; however, it creates a fuzzy cluster which incorporates observations that are far from fuzzy member centroids. These clusters are often areas with a high uncertainty of model predictions. The approach used here follows Malone et al. (2011), who used fuzzy membership to classify environmental covariates with a similar distribution of model errors. In each fuzzy cluster, a prediction interval is created by taking the weighted mean of the model errors. A detailed description of the process implemented can be found in Malone et al. (2017).

Results and discussion

Covariate importance

The five most important covariates for each soil property are shown in Table 6. The covariate importance for gravel, sand, clay and SOC is the scaled absolute value of the coefficients. In other words, the further from zero the coefficient, the higher the importance. The covariate importance for silt is based on the root squared error, or in other words, the larger the increase in the root squared error when the covariate is removed, the larger the importance of that covariate.

The number of covariates used to predict gravel, sand, silt, clay and SOC was 17, 12, 16, 19 and 14, respectively. For each soil property, topographic covariates were used in a larger proportion than spectral covariates. Sand and clay were the only properties that heavily relied on spectral covariates. Overall, the most important covariates were associated with slope position and shape, soil moisture and solar radiation. Therefore, topography is influencing soil distribution through erosion and depositional processes as well as solar radiation (Beaudette and O'Geen 2009; Brungard et al. 2015). In addition, a multitemporal approach is important when using spectral covariates. For example, a RI from February and a SI from August are among the top five important covariates for clay. These relationships would have been difficult to detect from expert knowledge alone and this justifies the feature selection approach.

Table 5: Algorithms used for all predictive models

Algorithm	Model type
General additive with splines (P-splines)	Generalised additive
Stochastic gradient boosting (SGB)	Tree-based additive
Ridge regression (RR)	Regularised linear
Linear boosted regression (LBM)	Linear additive
Quantile regression (QR)	Linear
Support vector machine (SVM)	Linear and radial
Cubist	Rule-based
Random forest (RF)	Tree-based

Predictive models

In this study, many predictive models were run because it cannot be assumed that a model will outperform another (Wolpert and Macready 1996; Kuhn and Johnson 2013). The strategy here was to try a number of models and focus on the top-performing model. The results for the top-performing feature selection and predictive model combinations are shown in Table 7. Ridge regression for both gravel and SOC had an optimal λ value of 0.01, which is the only tuning parameter. The sand SVM model had a linear kernel with a cost function of one. The silt RF model was optimised with 1 000 trees and 16 covariates randomly selected at each split. According to the out-of-bag error, the RF model explained 48% of the variance of silt, which corresponds well with the validation results before regression kriging. Clay predictions were best estimated by a linear quantile regression with no tuning parameters.

Satisfactory results were achieved for gravel and sand content in terms of RMSE and R^2 values. Both property predictions are comparable to other studies. For example, Ballabio et al. (2016) achieved an accuracy of 19.22% and an R^2 of 0.73 for gravel content. The authors used multivariate adaptive regression splines to map coarse fragments on the continental scale. In Burkina Faso, Forkuor et al. (2017) achieved an internal validation accuracy of 7.59% and R^2 of 0.34 for sand content using SVM. However, the authors achieved a more satisfactory result using multilinear regression. Silt, clay and SOC had less satisfactory R^2 values; however, the RMSE values were similar to other studies. In Kenya, Mutuma et al. (2016) used RF to predict silt and clay content with an accuracy of 7.30% and 9.90%, respectively. In Mozambique, Cambule et al. (2013) mapped SOC with an accuracy of 0.42% through kriging with external drift. It should be noted, however, that these studies were conducted at a regional, national or continental scale.

These results show a trend towards Boost feature selection with robust linear models. This suggests that this combination can be used as a powerful alternative to more complex models when predicting soil properties from a small data set. An explanation for gravel and SOC results is that the L_2 regularisation is suitable for small data sets (Kuhn and Johnson 2013) and slightly increases the bias to lower the variance (Hastie et al. 2009). For sand, the SVM with a linear kernel is suitable for high-dimensional data sets (Drucker et al. 1996). A quantile regression accounts for the skewed clay distribution and, therefore,

Table 6: Top five most important covariates for each soil property where rank, represents the order of importance. SOC = soil organic carbon

Rank	Gravel	Sand	Silt	Clay	SOC
1	Mid-slope position	Flow path length	Flow path length	RI fallow	Convexity
2	LSWIR ploughed	SWI	SWI	Total curvature	Normalised height
3	LS factor	Negative openness	Normalised height	Flow line curvature	SSWIR ploughed
4	Normalised height	Mid-slope position	Negative openness	Negative openness	Slope height
5	VRI	VRI	Convexity	SI growing	Aspect

quantile regression does not need normalised data (Koenker et al. 1978). Alternatively, these results may be a result of complex models over-fitting the small data set (Hastie et al. 2009).

Spatial autocorrelation

The sample variogram and variogram parameters for each soil property are shown in Figure 4. It should be noted that the soil properties did not need to be transformed as all residuals had a normal distribution according to a Shapiro normality test ($p < 0.05$). Spatial prediction of all the soil properties improved with regression kriging, but this improvement was relatively small. Gravel, sand, silt, clay and SOC accuracy (RMSE) improved by 0.11%, 0.15%, 0.17%, 0.14% and 0.007%, respectively. Perhaps the largest improvement can be seen in the R^2 values of clay and SOC, which improved by 7% and 5%, respectively. The sample variogram for gravel, sand and clay showed the most spatial autocorrelation and have the most reliable predictions.

The variogram for silt and SOC have a high nugget to sill ratio and may be unreliable. This can be attributed to the variograms being estimated from 93 soil observations, which is below the minimum sample size recommended by Oliver and Webster (2014). In addition, the lack of spatial autocorrelation could be due to the sample design (Webster and Oliver 2001). For example, over 75% of the soil samples are over 300 m apart with an average spacing of 763 m. Therefore, the processes determining the spatial distribution of SOC are acting on a scale that might not be represented on this site and/or from the soil sample design.

Table 7: Results for the top-performing model for each soil property. SOC = soil organic carbon, RMSE = root mean square error

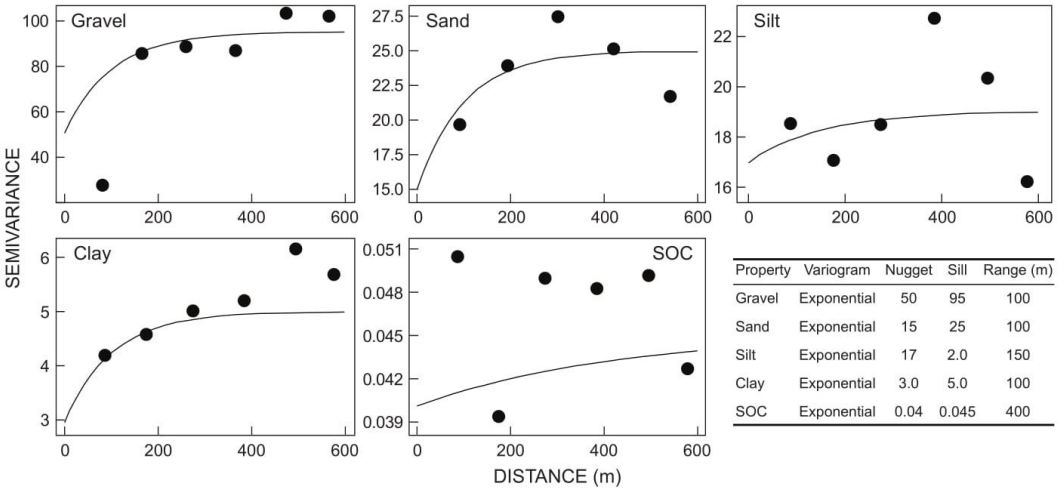
Property	Model	Krige	Selection	RMSE (%)	R^2
Gravel	RR	Yes	Boost	9.01	0.75
Sand	SVM	Yes	Boost	4.69	0.67
Silt	RF	Yes	RFS	4.12	0.53
Clay	QR	Yes	Boost	2.38	0.52
SOC	RR	Yes	Boost	0.19	0.41

Final predictions

Final predictions for all soil properties are shown in Figure 5. The final maps were discretised into three prediction quantiles. The legend for each soil property can be seen in Table 8. This was to simplify the maps to increase interpretability.

Gravel, silt and SOC spatial predictions appear realistic and mirror what was visually observed in the field. The initial maps for sand and clay were significantly affected by spectral images on different fields, causing discrete and unrealistic boundaries. These boundaries were seen after regression kriging. When spectral images were removed, the maps appeared more realistic. However, the prediction accuracy for sand decreased by 0.92% and the R^2 decreased by 14%. For clay, the accuracy decreased by 0.31% and the R^2 decreased by 11%. Due to the more realistic maps produced without spectral covariates, both sand and clay predictions had spectral covariates removed. These results suggest that the top-performing model may not be the optimal model to produce a farm-scale soil map. Therefore, each map should be inspected visually in addition to statistical evaluation. It is recommended that the simplest most realistic model be used as the final product.

Preferential erosion has removed the finer particles on crest and mid-slope positions resulting in the residual accumulation of gravel upslope (Shi and Schulin 2018). On the other hand, fluvial sands have resulted in the absolute accumulation of sand on lower elevations and, therefore,



Property	Variogram	Nugget	Sill	Range (m)
Gravel	Exponential	50	95	100
Sand	Exponential	15	25	100
Silt	Exponential	17	2.0	150
Clay	Exponential	3.0	5.0	100
SOC	Exponential	0.04	0.045	400

Figure 4: Residual variogram for all top-performing regression models

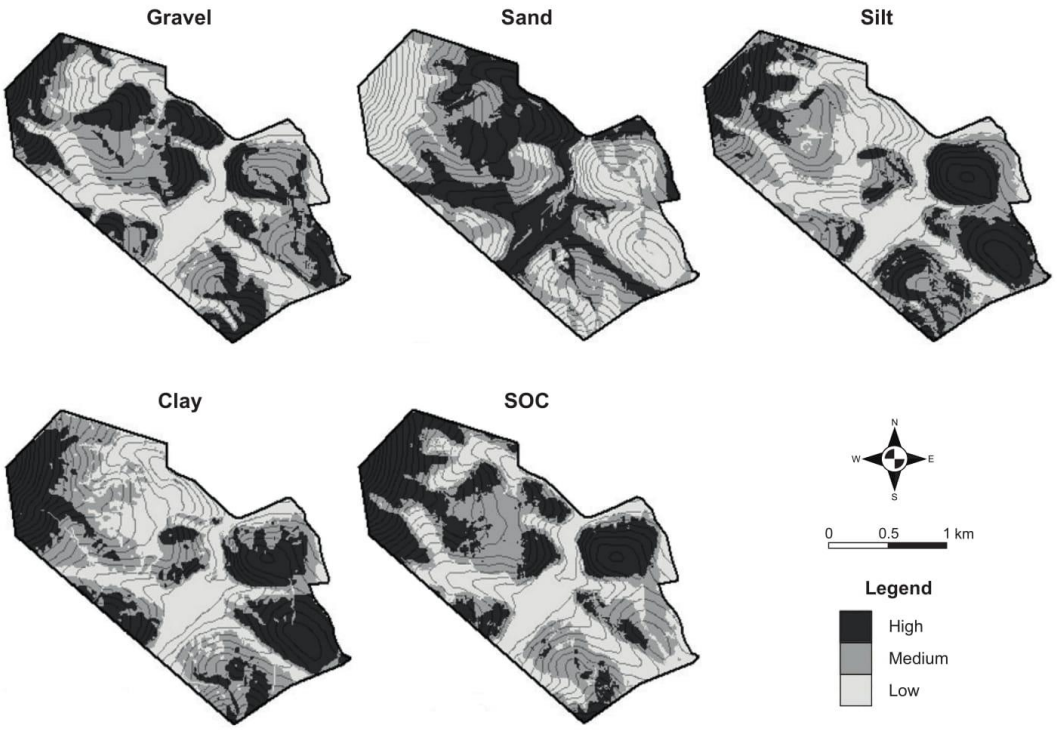


Figure 5: Prediction and prediction range for gravel, sand, silt, clay and soil organic carbon (SOC) content

Table 8: Legend for soil property prediction rasters

Property	Low		Moderate		High	
	Min%	Max%	Min%	Max%	Min%	Max%
Gravel	0.0	39.0	39.0	49.0	49.0	70.0
Sand	37.1	56.5	56.5	63.4	63.4	78.4
Silt	19.5	27.1	27.1	31.0	31.0	41.3
Clay	3.88	8.93	8.93	11.5	11.5	22.6
SOC	0.01	0.58	0.58	0.71	0.71	1.03

sand has a decreasing trend downslope. Soils at higher elevations developed from residual highly weathered material of the old African surface (Lambrechts 1983; Scholms et al. 1983). Soils developed from this parent material have a higher iron content than the younger soils below, which could be preventing the removal of finer particles by stabilising the soil aggregates (Barral et al. 1998). The higher clay content on upslope positions may also be protecting the SOC through sorption and micro-aggregation (Singh et al. 2018). Therefore, upslope positions have a higher silt, clay and SOC content.

Spatial uncertainties

To evaluate soil property uncertainties, prediction ranges were created through FKMe. Figure 6 and Table 9 show the prediction range and descriptive statistics for each soil property, respectively. The extragrades cluster for most of the soil properties correspond to what was reported by McBratney and De Gruijter (1992) as the extragrades cluster encompasses places of high uncertainty. However, the extragrades cluster for gravel has the lowest uncertainty. Furthermore, the extragrades cluster represents places where there is no soil, such as the stream or places with dense bush, which were not sampled. Therefore, care is recommended when interpreting this cluster. Besides the extragrades cluster, places of highest uncertainty are associated with concave slopes.

Conclusion

This study has shown that feature selection predictive model optimisation is effective at predicting gravel, sand, silt, clay and SOC at the farm-scale in the Sandspruit catchment. The models were evaluated on end accuracies for each individual soil property. The spatial uncertainties were evaluated through prediction ranges developed through FKMe. This approach is important where financial resources are low but high-resolution soil data is required. The conclusions of this study are:

- the simultaneous optimisation of feature selection and

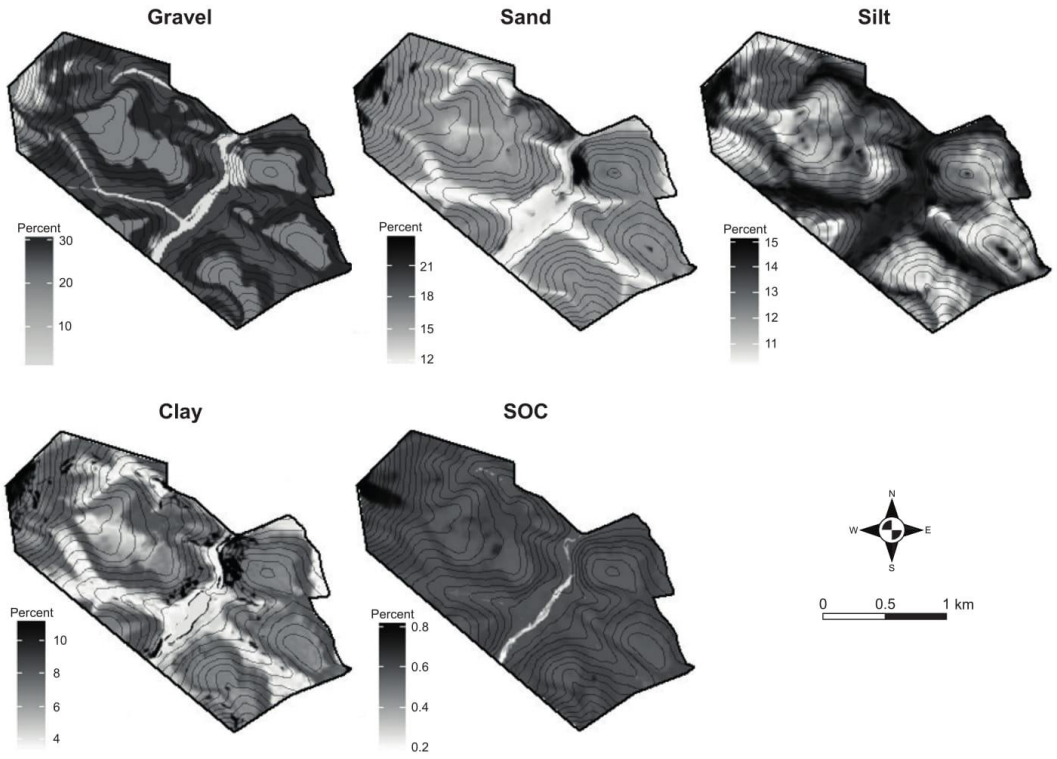


Figure 6: The range (%) for gravel, sand, silt, clay and soil organic carbon (SOC) predictions

Table 9: Descriptive statistics for prediction range of each soil property. SOC = soil organic carbon

Property	Clusters	Minimum (%)	Mean (%)	Maximum (%)
Gravel	4	0.0	25.0	30.0
Sand	3	11.6	15.8	23.7
Silt	3	10.2	12.8	15.1
Clay	3	3.40	6.74	11.3
SOC	3	0.00	0.63	1.06

predictive models proves to be a robust approach to predict soil properties at the farm-scale

- Boost feature selection and robust linear models obtained the highest end accuracy for four out of five soil properties
- regression kriging increased the accuracy for all soil property predictions
- spatial uncertainties suggest the highest uncertainty is associated with slopes (as opposed to hill tops and valleys).

This research lays out a DSM framework and, in theory, the methodology can be applied across South Africa at a detailed scale. A cost–benefit analysis is required to determine if this approach can lower the financial burden of traditional farm-scale soil surveys. In addition, this framework needs to be evaluated in different geographic areas and at different scales.

Geolocation information

The research site is bounded by latitudes 33°15′34.27″ S to 33°14′11.44″ S and longitudes 18°48′0.54″ E to 18°49′45.41″ E in the Swartland, Western Cape, South Africa.

Acknowledgements

This research was funded by the National Research Foundation of South Africa (reference SFH170525233469). This research was also assisted by Stellenbosch University. This work would not have been possible without the help of Dr Freddie Ellis who developed the sample design.

Disclosure statement

The authors declare no conflicts of interest.

References

Aitkenhead MJ, Coull MC. 2016. Mapping soil carbon stocks across Scotland using a neural network model. *Geoderma* 262: 187–198.

Ballabio C, Panagos P, Monatanarella L. 2016. Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma* 261: 110–123.
Barral MT, Arias M, Guerif J. 1998. Effects of iron and organic matter on the porosity and structural stability of soil aggregates. *Soil and Tillage Research* 46: 261–272.
Beaudette DE, O'Geen AT. 2009. Quantifying the aspect effect: an application of solar radiation modeling for soil survey. *Soil Science Society of America Journal* 73: 1345–1352.
Behrens T, Förster H. 2005. Digital soil mapping using artificial neural networks. *Journal of Plant Nutrition and Soil Science* 25: 580–591.
Behrens T, Zhu AX, Schmidt K, Scholten T. 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma* 155: 175–185.
Bishop TFA, McBratney AB. 2001. A comparison of prediction methods for the creation of field-extent soil property maps. *Geoderma* 103: 149–160.
Brungard CW, Boettinger JL, Duniway MC, Wills SA, Edwards TC. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma* 239–240: 68–83.
Bui EN, Moran CJ. 2000. Regional-scale investigation of the spatial distribution and origin of soluble salts in central north Queensland. *Hydrological Processes* 14: 237–250.
Cambule AH, Rossiter DG, Stoorvogel JJ. 2013. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192: 341–353.
Chagas CS, Carvalho Junior W, Bhering SB, Calderano Filho B. 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena* 139: 232–240.
De Gruijter JJ, Walvoort DJJ, Van Gaans PFM. 1997. Continuous soil maps — a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma* 77: 169–195.
Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. 1996. Support vector regression machines. In: Mozer MC, Jordan MI (eds), *Advances in neural information processing systems 9: proceedings of the 1996 conference*. Cambridge, MA: MIT Press. pp 155–161.
Forkuor G, Hounkpatin OKL, Welp G, Thiel M. 2017. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PLOS ONE* 12(1): e0170478.
Hastie T, Tibshirani R, Friedman J. 2009. *The elements of statistical learning* (2nd edn). *Springer Series in Statistics*. New York: Springer.
Hengl T. 2006. Finding the right pixel size. *Computers and Geosciences* 32: 1283–1298.
Hengl T, Reuter HI (eds). 2009. *Geomorphology: concepts, software, applications*. *Developments in Soil Science* vol. 33. Amsterdam: Elsevier.
Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD, Sila A, MacMillan RA, Jesus JM, Tamene L, Tondoh JE. 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. *PLoS ONE* 10: e0125814.
Hoffmann U, Hoffmann T, Jurasinski G, Glatzel S, Kuhn NJ. 2014. Assessing the spatial variability of soil organic carbon stocks in an alpine setting (Grindelwald, Swiss Alps). *Geoderma* 232–234: 270–283.
Huete AR. 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment* 25: 295–309.
Jafari A, Khademi H, Finke PA, Van de Wauw J, Ayoubi S. 2014. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. *Geoderma* 232–234: 148–163.
Kerry R, Oliver MA. 2007. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma* 140: 383–396.
Koenker R, Bassett G, Jan N. 1978. Regression quantiles. *Econometrica* 46: 33–50.
Kuhn M, Johnson K. 2013. *Applied predictive modeling*. New York: Springer.
Lagacherie P. 1992. Formalisation des lois de distribution des sols pour automiser la cartographie pedologique a partir d'un secteur pris comme reference. Universite Montpellier II, France.
Lagacherie P, Cazemier DR, Van Gaans PFM, Burrough PA. 1997. Fuzzy *k*-means clustering of fields in an elementary catchment and extrapolation to a larger area. *Geoderma* 77: 197–216.
Lambrechts JJN. 1983. *Soils, soil process and distribution in the Fynbos region: an introduction*. In: Deacon HJ, Hendey QB, Lambrechts JJN (eds), *Fynbos palaeoecology: a preliminary synthesis. South African National Scientific Programmes Report* no. 75. Pretoria: Council for Scientific and Industrial Research. pp 61–69.
Lark RM, Cullis BR, Welham SJ. 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science* 97: 787–799.
Malone BP, McBratney AB, Minasny B. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* 160: 614–626.
Malone BP, Minasny B, McBratney AB. 2017. *Using R for digital soil mapping*. Cham: Springer International Publishing.
Mason CH, Perreault WD. 1991. Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research* 28: 268–280.
McBratney AB, De Gruijter JJ. 1992. A continuum approach to soil classification by modified fuzzy *k*-means with extragrades. *Journal of Soil Science* 43: 159–175.
McBratney AB, Santos MLM, Minasnya B. 2003. On digital soil mapping. *Geoderma* 117: 3–52.
McBratney AB, Webster R. 1983. Optimal interpolation and isarithmic mapping of soil properties: V. Co-regionalization and multiple sampling strategy. *European Journal of Soil Science* 34: 137–162.
McKenzie NJ, Austin MP. 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. *Geoderma* 57: 329–355.
Minasny B, McBratney AB. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences* 32: 1378–1388.
Minasny B, McBratney AB. 2016. Digital soil mapping: a brief history and some lessons. *Geoderma* 264: 301–311.
Moran CJ, Bui EN. 2002. Spatial data mining for enhanced soil map modelling. *International Journal of Geographical Information Science* 16: 533–549.
Mora-Vallejo A, Claessens L, Stoorvogel J, Heuvelink GBM. 2008. Small scale digital soil mapping in southeastern Kenya. *Catena* 76: 44–53.
Mutuma E, Csorba A, Michéli E. 2016. Prediction of soil properties using Mid-Infrared Spectroscopy and Random Forest regression in the Eastern slopes of Mt. Kenya Region. *Agricultural Science Research Journal* 6: 253–262.
Oliver MA, Webster R. 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena* 113: 56–69.
Pahlavan Rad MR, Toomanian N, Khormali F, Brungard CW, Komaki CB, Bogaert P. 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. *Geoderma* 232–234: 97–106.

Qi F, Zhu AX, Harrower M, Burt JE. 2006. Fuzzy soil mapping based on prototype category theory. *Geoderma* 136: 774–787.
Ray SS, Singh JP, Das G, Panigrahy S. 2004. Use of high resolution remote sensing data for generating site-specific soil mangement plan. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 35(B7): 127–131.
Scholms BHA, Ellis F, Lambrechts JJN. 1983. Soils of the Cape Coastal Platform. In: Deacon HJ, Hendey QB, Lambrechts JJN (eds), *Fynbos palaeoecology: a preliminary synthesis. South African National Scientific Programmes Report* no. 75. Pretoria: Council for Scientific and Industrial Research. pp 70–86.
Shi P, Schulin R. 2018. Erosion-induced losses of carbon, nitrogen, phosphorus and heavy metals from agricultural soils of contrasting organic matter management. *Science of the Total Environment* 618: 210–218.
Singh M, Sarkar B, Sarkar S, Churchman J, Bolan N, Mandal S, Menon M, Purakayastha TJ, Beerling DJ. 2018. Stabilization of soil organic carbon as influenced by clay mineralogy. *Advances in Agronomy* 148: 33–84.
Soil Survey Staff. 2014. *Keys to soil taxonomy* (12th edn). Washington, DC: US Department of Agriculture, Natural Resources Conservation Service. Available at http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_051546.pdf.
Soil Survey Staff. 2017. *Soil survey manual* (4th edn). *Agriculture Handbook* no. 18. Washington, DC: US Department of Agriculture, Natural Resources Conservation Service, Soil Science Division.
Subburayalu SK, Jenhani I, Slater BK. 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. *Geoderma* 213: 334–345.
Sun XL, Zhao YG, Zhang GL, Wu SC, Man YB, Wong MH. 2011. Application of a digital soil mapping method in producing soil orders on mountain areas of Hong Kong based on legacy soil data. *Pedosphere* 21: 339–350.
Van Zijl GM, Le Roux PA, Turner DP. 2013. Disaggregation of land types using terrain analysis, expert knowledge and GIS methods. *South African Journal of Plant and Soil* 30: 123–129.
Walkley A, Black IA. 1934. An examination of Degtjareff method for determining soil organic matter, and proposed modification of the chromic acid tritation method. *Soil Science* 37: 29–38.
Webster R, Oliver MA. 2001. *Geostatistics for environmental scientists* (2nd edn). Chichester: Wiley.
Wolpert DH, Macready WG. 1996. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* 1: 67–82.
Xu Y, Smith SE, Grunwald S, Abd-Elrahman A, Wani SP, Nair VD. 2018. Estimating soil total nitrogen in smallholder farm settings using remote sensing spectral indices and regression kriging. *Catena* 163: 111–122.
Zhu AX. 1997. A similarity model for representing soil spatial information. *Geoderma* 77: 217–242.

Received 10 September 2018, revised 6 January 2019, accepted 6 January 2019
Associate Editor: Johan van Tol

Supplementary Information

High-resolution digital soil mapping of multiple soil properties: an alternative to the traditional field survey?
Trevan Flynn, Willem de Clercq, Andrei Rozanov and Cathy Clarke
South African Journal of Plant and Soil 2019. https://doi.org/10.1080/02571862.2019.1570566

Appendix S1: Feature selection implementation

All covariates were centred and scaled before running the feature selection. The univariate feature selection (UFS) fits a model for many iterations after filtering of the covariates. Covariate filtering selects covariates by determining the correlation of individual covariates to the soil property of interest. The robust feature selection (RFS) takes all the covariates and progressively eliminates each until the error rate reaches an optimal level. No tuning parameters were optimized for either technique.

The least absolute shrinkage and selection operator (LASSO) is a generalized linear model which minimizes covariate coefficients based on the absolute error of the residuals (L_1 regularisation) through coordinate descent (Friedman et al., 2010). This process shrinks covariate coefficients which are correlated to one another. The degree of shrinkage is controlled by the λ value which was optimised and the covariates which did not have an absolute value of zero, were selected. A LASSO feature selection was implemented because LASSO is efficient with high dimensional data sets, improves model interpretation, and does not substantially increase bias (Tibshirani, 1996).

A boosted linear model (Boost) is a novel feature selection technique suitable for high dimensional data sets (Bühlmann & Hothorn, 2007). It fits component-wise linear models as base learners and is boosted by correcting for the squared error of the residuals (L_2 regularisation). However, unlike L_1 regularisation, the coefficients are not shrunk to zero and the method of feature selection is a “black box” with little known as to how it selects the covariates. The number of boosts was optimised with pruning.

Predictive model implementation

A RR is a generalized linear model which maximises the likelihood via L_2 regularisation through coordinate descent (Friedman et al., 2010). The λ value was the only parameter optimised. An LBM is an additive model with linear step-wise base learners.

The number of boosts were optimised with pruning. The QR was implemented through regression on the median with no tuning parameters.

Both linear and radial kernel SVM were implemented to evaluate both linear and non-linear relationships. Support Vector Machines have been known to perform well for classification, however, these have been adapted to perform regression tasks. The cost function and sigma values were both optimised. The cost controls the error function of the model. The sigma value determines the width of the gaussian distribution for the radial kernel.

Random forest is a decision tree ensemble model which grows trees in parallel and the final prediction is the mean of the prediction for all trees grown (Breiman, 2001). The number of covariates randomly chosen at each split was optimised and the number of trees grown was held constant at 1000 trees. The number of trees was held at 1000 because Breiman (2002), states that at least 1000 trees are required for a stable variable importance measure. Random forest was used because it is suitable for small and large data, can handle non-linear relationships, and is robust against over fitting (Breiman, 2001).

Stochastic gradient boosting is a type of decision tree ensemble which creates decision trees in sequence rather than parallel (Friedman, 2001, 2002). Therefore, the model builds decision trees to correct for the errors of the previous decision tree. The SGB algorithm implements a gaussian exponential loss function through Friedman gradient decent (Friedman, 2001). The learning rate, minimum number of observations in each terminal node, and the bag fraction were held constant at 0.01, ten observations, and 0.5 resamples, respectively. However, the number of trees grown and number of interactions was optimised. A SGB was used because it represents an alternative to RF and has been shown to achieve similar accuracies (Forkuor et al., 2014).

The Cubist algorithm is a rule based model which runs linear regression as a smoothing parameter (Quinlan, 1993). The cubist model is similar to an ensemble of decision trees; however, linear regression is performed at each node. The cubist model has two main tuning parameters. The number of committees is the number of trees grown in sequence (like boosting). The number of neighbours is the number of k-nearest neighbours used to correct for errors. The cubist model was selected

because cubist is a complex yet an interpretable model as the output defines each rule made.

Penalized boosted splines is an additive model which uses splines as a smoothing base learner for each covariate and the model is boosted on the residuals (Bühlmann & Hothorn, 2007). The number of boosts was optimised with pruning, knots were set to 20, and degrees of freedom set to four. The knots and degrees of freedom values were held constant based on the recommendations of Bühlmann and Hothorn (2007). Penalized boosted splines were implemented because it is a novel algorithm which has been shown to be a powerful tool in machine learning competitions (Taieb & Hyndman, 2013).

References

Breiman L. 2001. Random forests. *Machine Learning* 45(1): 5–32.

Breiman, L. 2002. Manual on setting up, using, and understanding random forests v3.1. Technical Report. Available at http://oz.berkeley.edu/users/breiman, Statistics Department University of California Berkeley.

Bühlmann, P. & Hothorn, T. 2007. Boosting algorithms: regularization, prediction and model fitting. *Statistical Science* 22(4): 477–505.

Forkuor, G., Hounkpatin, O.K.L., Welp, G. & Thiel, M. 2017. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PLOS One* 12(1): e0170478.

Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29(5): 1189–1232.

Friedman, J.H. 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis* 38: 367–378.

Friedman, J., Hastie, T. & Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1): 1–22.

.

Hitziger, M. & Ließ, M. 2014. Comparison of three supervised learning methods for digital soil mapping: application to a complex terrain in the Ecuadorian Andes. *Applied and Environmental Soil Science* 2014: Art. ID 809495.

Quinlan, J.R. 1993. Combining instance-based and model-based learning. *Machine Learning* 76: 236–243.

Taieb, S. Ben & Hyndman, R.J. 2013. A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting* 30(2): 382–394.

Tibshirani, R. 1996. Regression selection and shrinkage via the Lasso. *Journal of the Royal Statistical Society* 58(1): 267–288.

Appendix B

Geoderma 337 (2019) 1136–1145

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Semi-automatic disaggregation of a national resource inventory into a farm-scale soil depth class map

Trevan Flynn[a], Andrei Rozanov[a], Willem de Clercq[b], Benjamin Warr[c], Cathy Clarke[a,*]

[a] Department of Soil Science, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
[b] Stellenbosch Water Institute, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
[c] BetterWorld Energy Ltd. Chitemwiko Close, Kabulonga, Lusaka, Zambia

ARTICLE INFO

Editor: Alex McBratney

Keywords:
Digital soil mapping
DSMART
Farm-scale
Geomorphons

ABSTRACT

Knowledge of soil depth spatial variability is important for land use management especially in dryland agriculture regions, which rely on climate and soils to provide adequate water and nutrients during the growing season. Soil spatial variability can be predicted from legacy soil data through machine learning techniques producing quantitative soil maps requiring minimal resources. South Africa has a country wide 1:250,000 scale resource map known as the Land Type Survey (LTS) which includes soil properties such as soil depth, soil class, root limiting layer, clay content, and texture. Each LTS polygon (land type), is comprised of unique soil – terrain patterns and is therefore, not a true soil map. This study aims to disaggregate the LTS into a farm-scale soil depth class map through a two-step disaggregation approach. First, landform elements were predicted through a pattern recognition algorithm known as geomorphons. Geomorphons, together with the original LTS were overlaid to produce polygons with unique distributions of soil. The polygons were disaggregated further to produce a raster map of soil depth classes through a soil map disaggregation algorithm known as DSMART. The first most probable class raster achieved an accuracy of 68% and for the two most probable class rasters, an accuracy of 91% was achieved. The two-step approach proved necessary for producing a farm-scale soil map. The result of this study is significant as it produced a soil depth class map from a national resource map at a scale and resolution (10 m) suitable for farm management.

1. Introduction

Soil depth is one of the most important properties of soils and is one of the seven major considerations when evaluating soil quality (Bunning et al., 2011). Soil depth and its spatial variability is crucially important for activities such as planning irrigation (Myburgh et al., 1996), hydrological modelling (Devia et al., 2015), estimating soil carbon stocks (Wiese et al., 2016), and many other soil management practices. Soil properties, including depth, may be highly variable on old land surfaces due to multiple cycles of erosion and deposition (Rozanov et al., 2017). Soil depth is difficult to estimate, and usually requires physical observations which are expensive. Alternatively, soil depth can be predicted through digital soil mapping (DSM) which requires fewer soil samples and, when soil legacy data is available, no soil samples at all. As increasingly high-resolution ancillary data becomes available, DSM methods such as the *scorpan* method (McBratney et al., 2003), can be adapted to suit local needs and make better use of legacy data, for instance, by generating farm-scale (1:20,000) maps at an appropriate resolution.

South Africa has a 1:250,000 scale land resource map known as the LTS, which took over 30 years to complete and involved > 20 soil scientists (Land Type Staff, 1972–2002). Although not a true soil map, the LTS gives information to inform agricultural potential of South Africa by mapping unique distributions of climate, terrain, and soil. The soil information contained in the LTS includes local soil classes (soil series), soil depth ranges, root limiting layer, clay percent, and texture. Due to the nationwide coverage and the amount of soil information contained within, van Zijl et al. (2013) states that DSM in South Africa should begin with the disaggregation the LTS.

The LTS was developed in a hierarchical structure that is spatially scaled from the top down. At the top, are climatic regions which divide the country. Within each climatic region, areas are dissected by uniform patterns of landform elements (LFEs) and repeating soil patterns which are restricted to the individual polygons on the 1:250,000 map. These polygons are known as land types and each consist of a unique legend. The LTS legend displays LFEs as the percent area of each land type and

* Corresponding author.
E-mail address: cdowding@sun.ac.za (C. Clarke).

https://doi.org/10.1016/j.geoderma.2018.11.003
Received 6 May 2018; Received in revised form 28 October 2018; Accepted 3 November 2018

soil information is displayed as percent area of each LFE within each land type (Land Type Survey Staff, 1972-2002). However, both LFEs and soil information were estimated from a recognisance scale soil survey and were not externally validated.

There are many difficulties facing the disaggregation of the LTS, such as the coarse spatial extent, coarse feature scale, and the fact that the LTS does not have true soil polygons. The coarse spatial extent and feature scale of the LTS was necessary for nationwide coverage; however, this information is too coarse for land use management. To generate finer-scale soil information, LFEs must be classified to disaggregate the coarse land type polygons into basic LFEs that have associated soil attributes. In doing so, the LTS can be stratified into disaggregated polygons with more homogenous distribution of soil. Therefore, each disaggregated polygon has a unique distribution of soil consisting of either unique soil attributes or more commonly, having the same soil attributes with different probabilities. There are many LFEs classification algorithms which include nested-clustering methods (Iwahashi and Pike, 2007), fuzzy landform classification (Irvin et al., 1997; Schmidt and Hewitt, 2004), object-based classification (Dragut, 2011), and more. However, these algorithms are not flexible to changes in scale, are often computationally inefficient, and/or do not coincide with slope position used by soil scientists in delineating the original LFEs (Jasiewicz and Stepinski, 2013; Libohova et al., 2016).

Developed by Jasiewicz and Stepinski (2013), geomorphons is a pattern recognition algorithm created to be computationally efficient, flexible to scale, and can correspond to slope position. Geomorphons classify the ten most common LFEs known as flat, peak, ridge, shoulder, spur, slope, hollow, foot-slope, valley, and pit. Silva et al. (2016) demonstrated the flexibility of the geomorphon approach by stratifying a tropical soil landscape to help soil surveying in two different catchments in Brazil. Libohova et al. (2016) showed that by aggregating 10-unit geomorphon (GM-10) into a 5-unit geomorphon (GM-5) that correspond to slope position such as foot-slope and toe-slope, LFEs could be predicted with an 81% accuracy from soil properties.

Other factors making the disaggregation of the LTS difficult is that there are very few georeferenced soil profiles and many land types do not have any georeferenced soil profiles. The lack of soil profiles makes updating the LTS problematic without an additional soil survey. Additionally, the LTS was developed by different soil scientists who surveyed different land types and the understanding of how to delineate the land types developed through its creation. Therefore, different soil environmental rules were established depending on who the surveyor was and the time of the survey.

There are many approaches that have been developed that incorporate existing resource inventories into the DSM framework (Grunwald, 2006; Minasny and McBratney, 2016; Scull et al., 2003). These techniques include geostatistics, expert knowledge systems, and machine learning algorithms which have been successfully applied through kriging with external drift (Kempen et al., 2015), fuzzy logic (MacMillan et al., 2000; Silva et al., 2014; Smith et al., 2010; Yang et al., 2011), k-means clustering (Bui and Moran, 2001), decision trees (Nauman and Thompson, 2014; Sarmento et al., 2017; Subburayalu et al., 2014), and random forest (Häring et al., 2012; Nauman et al., 2014). An expert knowledge system through a fuzzy logic inference system known as SoLIM (Zhu, 1997), has been successfully applied to update the LTS into soil associations by van Zijl et al. (2013). However, these techniques often require additional soil samples, are restricted to soil polygon boundaries, and/or predict individual soil classes separately.

A technique known as “Disaggregating and Harmonisating of Soil Map Units Through Resampled Classification Trees” (DSMART) developed by Odgers et al. (2014) shows promise to disaggregate the LTS into a soil map. The DSMART algorithm uses resampled classification trees to create multiple realisations from soil legacy polygons. For each realisation, DSMART finds soil environmental relationships through the randomly assigned samples and covariates. These realisations are used to calculate the probability of each soil class and make the final predictions of a specified number of probable class rasters. The algorithm can be seen as a stochastic approach where multiple realisations give an estimate of the soil distribution. A benefit of the DSMART algorithm is that it can predict soil classes across soil polygon boundaries and predict all soil classes simultaneously. This would be beneficial to disaggregate the LTS where soil information needs to be predicted across boundaries created by different surveyors and LFEs. Additionally, DSMART does not necessarily need to be implemented with classification trees, making it flexible to algorithm chosen and soil attribute to be predicted.

Odgers et al. (2014) who developed DSMART, showed DSMART's potential by disaggregating a soil legacy map in central Queensland, Australia. The authors achieved a 22.5% accuracy for the first most probable class raster. However, the three most probable class rasters combined, classified 50% of the soil classes correctly. Vincent et al. (2018) implemented DSMART to disaggregate a French resource map at a 1:250,000 scale. The authors achieved an accuracy of 41% to 72% depending on the validation technique used. Holmes et al. (2015) disaggregated a soil – terrain polygon map through DSMART to predict Soil Groups of Western Australia. The authors achieved an accuracy of 41% for the three most probable class rasters. Perhaps the most remarkable implementation of DSMART came from Chaney et al. (2016) who implemented DSMART with the Random Forest algorithm to disaggregate the Soil Survey Geographic database for the whole contiguous United States. The result of this study was a continuous soil series database known as POLARIS, which according to the first ten realisations, matched 55% of the Soil Survey Geographic database. These studies all focused on regional or even country scale and the algorithm has not yet been tested for farm-scale mapping (Malone et al., 2017).

This research serves as an investigative study to disaggregate the LTS into a farm-scale soil map at a resolution suitable for farming. Due to the importance of soil depth, this is the property selected for mapping. The process follows the hierarchical structure of the LTS by first stratifying the land types into homogenous distributions of soil, and then disaggregating these newly formed polygons through machine learning. The main idea behind this implementation was to adapt a large spatial scale DSM framework and adjust it for the local needs in South Africa. This can potentially increase the accessibility and usefulness of the LTS information to farmers by quantifying the spatial variability within the LTS.

2. Methods and materials

2.1. Site description

The research site lies in the middle of the Sandspruit catchment which forms a tributary to the Berg River in the Swartland district of the Western Cape Province, South Africa approximately 50 km north of Cape Town. The site location and the predominant land types in the Sandspruit catchment are shown in Fig. 1a and b respectively. The site encompasses 366 ha located approximately 33°14′53.34″ S and 18°48′53.32″ E. A small area was established to determine if the LTS can be disaggregated and downscaled into a larger scale map from the original 1:250,000 LTS map. The area is under dryland agriculture with wheat and canola crop rotation. The average rainfall in the region is 394 mm per year with winter rainfall. Temperature averages range from 16.2 °C in July to 27.4 °C in February. Altitude ranges from 94 m to 218 m above sea level. The site has a land surface of differing ages with the old African erosional surface at high elevation which is dissected by a younger surface at lower altitudes. The site was selected because of its similarity to the rest of the catchment and to capture as much terrain variability as possible within a reasonably small geographical area. Furthermore, the site was selected due to the multiple land types that intersect the site.

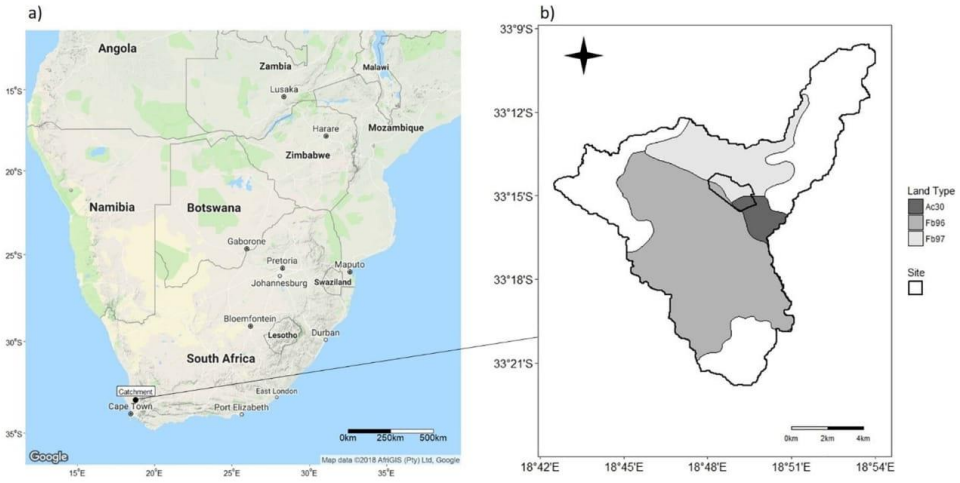


Fig. 1. The location of the Sandspruit catchment within South Africa (a) and the location of the site within dominant land types of the catchment (b).

There are three land types that dissect the site according to the Land Type Survey Staff (1972-2002). These land types are Ac30, Fb96, and Fb97 as displayed on the land type sheet 3318 (Cape Town). The predominant geology of the catchment is greywacke, phyllite and schist of the Moorreesburg, Klipplaat, and Berg River formations. There are also silcrete and ferricrete outcroppings scattered through the catchment. It is hypothesized that these outcroppings represent the transition from the old African erosion surface to the younger dissected surface below. There are two dominant soils in the catchment according the LTS. These are the relatively shallow residual soils such as Lithic haploxerepts and the deeper red apedal soils such as Typic haploxerept. These soils represent different parent materials and different ages. The residual soils are weathered from the shale parent material and are overlain by a thin, loamy, usually gravely creep layer. These residual soils are younger and are found at lower altitudes than the highly weathered drift from the old African surface which occurs on pre-weathered shale. The red apedal soils, form within this highly weathered drift material and are found at higher altitudes.

2.2. Data acquisition and pre-processing

2.2.1. Software

Geomorphons and covariates were developed from a digital elevation model (DEM) using the Geographic Resource Analysis Support System (GRASS v7.2; GRASS Development Team, 2017) and the System for Automated Geoscientific Analyses (SAGA v2.3.2; Conrad et al., 2015), respectively. All models and statistical analysis were conducted in R software (R Core Team, 2017). Disaggregation was developed from the DSMART algorithm as in the rdsmart package (Odgers and Malone, 2017). However, the rdsmart code was adapted to incorporate any classification model available in the caret R package (Kuhn et al., 2018).

2.2.2. Land Type Survey database

The land types were obtained from the LTS sheet 3318 (Cape Town) which has a basic map unit of 160 ha. The legends for each land type were gathered from separate text files. A depiction of the structure of the LTS data is shown in Table 1. The legends specify the LFEs found on each land type and the soil distributions on each LFE. The LFE classification used in the LTS is a 2-dimensional manually delineated classification which has five units that generalize places of convergence and divergence of water movement. A cross section of a land type's LFEs on this site (Fb96) is shown in Fig. 2. These LFEs are crest (1), scarp (2), mid-slope (3), foot-slope (4), and valley (5). The soil distributions defined in the legend consists of a taxonomic soil class (soil series) and the properties associated with each soil series such as soil depth ranges. To extract soil depth classes, the LTS depth ranges were averaged and divided into shallow (0 cm–40 cm), moderate (40 cm–80 cm) and deep classes (> 80 cm).

Table 1
The hierarchical structure of the data obtained for the LTS information, how the information is represented and how the files are obtained.

Information	Representation	File type
Land types	Spatial polygons	Shapefile
LFEs	Probabilities in each polygon	Legend
Soil attributes	Probabilities in each LFE	Legend

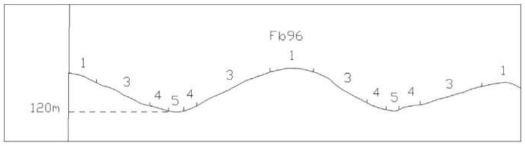


Fig. 2. Cross section of land type Fb96 showing LFEs crest (1), mid-slope (3), foot-slope (4), and valley (5).

2.2.3. Digital elevation model

A 30 m DEM was acquired from the Advanced Land Observation Satellite (ALOS-2) provided by the Japanese Aerospace Exploration Agency (JAXA) http://www.eorc.jaxa.jp/ALOS/en/aw3d30. The DEM was obtained from the S034E018 thumbnail corresponding to the geographic coordinates of 34° south 18° east. The DEM was selected because the vertical accuracy (6 m) is significantly higher than other freely available DEMs (i.e. 13 m SRTM) at the same resolution. The DEM was re-projected into the Hartebeeshoek94 datum projected coordinate system. Reprojection was necessary because the data sets were distributed on different coordinates systems, accurate distances between points was needed, and to locate the soil observations in the field. However, this can add distortion to the DEM and lower its

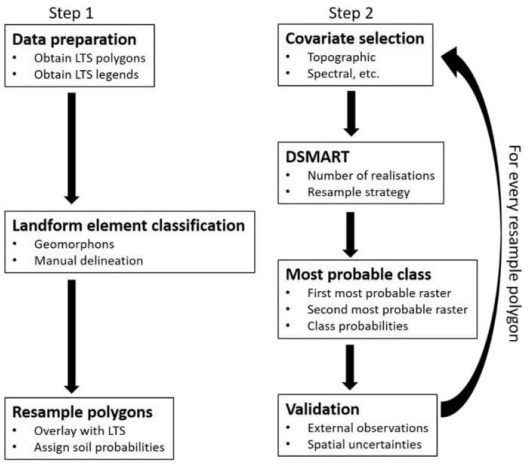


Fig. 3. Methodology for the disaggregation of the LTS by stratifying the LTS through geomorphons and running DSMART to extract the spatial distribution of the soil depth classes.

accuracy (Hengl and Evans, 2009). The DEM was resampled to a 10 m resolution and had sinks filled. A 10 m resolution is the finest legible resolution at the scale of this site (366 ha) according to inspection density (Hengl, 2006). The concept of inspection density is that maps are predicted from point data and therefore, should have a similar sample density per area. The equation for the finest legible resolution can be seen in Eq. (1). Where A is the area of the site in square meters (3,660,000 m^2) and N is the number of soil observations (93 observations).

$$\text{Finest Legible Resolution} = 0.05 \times \sqrt{\frac{A}{N}} \quad (1)$$

2.3. Disaggregation approach

The disaggregation approach in this paper follows a two-step method to disaggregate the LTS shown in Fig. 3. The first step of the processes was to classify LFEs corresponding to the LTS legend. The LFEs were overlaid with the LTS and soil depth class probabilities (percent area specified in the LTS) were manually assigned to each polygon. This created polygons (resample polygons) with unique distributions of soil. The resample polygons can be seen as soil legacy data stratified by LFEs thereby, creating polygons with a more detailed spatial scale than the original LTS. DSMART was run by drawing random samples from the resample polygons with a pool of covariates to establish soil environmental relationships to predict soil depth classes. The final product consisted of the first and second most probable class rasters, the probabilities of each soil class, and spatial uncertainties. Spatial uncertainties are an evaluation of how certain the model is of the predictions made through the extent of the area. This process was compared to resample polygons created from a manually delineated LFEs and using the original LTS as resample polygons with no LFE classification.

2.3.1. Landform element development

Geomorphons were used to classify LFEs because they are computationally efficient, account for landscape position, and are robust against scale (Jasiewicz and Stepinski, 2013). The algorithm utilizes a "line-of-sight" approach, relating two different angles over eight directions from a central point. The two angles are known as the zenith and nadir angles. The zenith angle is the angle between 90° (overhead) and the line of sight (0°). The nadir angle is the angle from −90° (below) to the line of sight. The user can specify two parameters in the algorithm, the search radius and flatness threshold. The search radius is the radius the algorithm will search away from a central point to distinguish landscape patterns. The flatness threshold defines what is and is not considered flat. A graphical representation of the LFE classified through the geomorphon algorithm can be seen in Fig. 4.

A geomorphon was developed through the i.geomorphons GRASS add-on developed by Jasiewicz and Stepinski (2013). The final geomorphon selected had a 200-cell search radius and a flatness threshold of 1°. It was thought that this geomorphon represented the rolling topography best and was most suitable to vectorise into polygons. To correspond with the LTS, the GM-10 was aggregated into a GM-5. It should be noted that the LTS has up to five LFEs, however, not all land types have all five LFEs. For example, the LTS legends for this site are classified as crest, mid-slope, foot-slope, and valley. Therefore, geomorphon aggregation needs to correspond to the legend for each individual land type. To aggregate the geomorphon units, the geomorphon values were changed for each individual pixel of the original GM-10. Peak, ridge, and shoulder were aggregated into crest as these positions are convex, usually high elevation positions with slopes accruing below. Spur and slope are generally found on mid-elevation positions and were therefore, aggregated into mid-slope. Hollow and foot-slope positions, which correspond to concave slopes, were aggregated into foot-slopes. Flat slopes were only found in valley positions and along with valley and pit, were aggregated into valley. The reclassification method used in this study is a simplification of the method developed by Libohova et al. (2016), who aggregated GM-10 into GM-5 based on a slope gradient threshold.

To determine if the geomorphon algorithm can mimic the "mental model" used to define LFEs in the LTS, an additional manually delineated LFEs was created (expert GM). The expert GM was developed by Dr. Freddie Ellis, a local pedologist with 50 years mapping experience in the region. Dr. Ellis was the surveyor for the LTS sheet 3318 (Cape Town) that covers the study area. The expert GM was created by delineating boundaries from a Google Earth satellite image overlaid with 5 m contour lines. The delineation of the expert GM was based on elevation, slope curvatures, and landscape position.

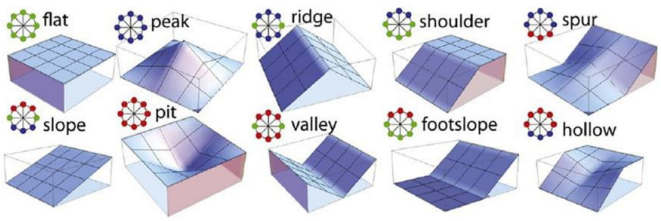


Fig. 4. The ten most common landform elements classified by geomorphons (Jasiewicz and Stepinski, 2013).

2.3.2. Resample polygons

Resample polygons were created for both the GM-5 and the expert GM by first converting the rasters into polygons and then overlaying with the land types. These resample polygons will be referred to as LTS-GM5 and the LTS-EX5, respectively. In the overlaying process, the probabilities of soil depth classes were manually assigned to each resample polygon. This created soil depth class probabilities according to the LTS for each LFE within each land type. The resample polygons can be seen as the first step in disaggregating the LTS as the LFEs disaggregate the land types from soil-terrain patterns, into soil patterns. The comparison between the two resample polygons is important as it is thought that the LTS-GM5 must mimic the LFEs generated by the LTS surveyor to obtain accurate results.

To determine if the two-step disaggregation approach is necessary, the LTS was disaggregated without LFE classification in a single step approach. This was done by taking the total soil depth class probabilities of each LTS legend. Therefore, in this implementation, the land types themselves were used as resample polygons. These resample polygons will be referred to as LTU. This comparison is important as the two-step disaggregation approach requires a greater in-depth knowledge of soil environmental relationships and since it requires two models, the two-step approach may have errors which propagate through each step.

2.3.3. Covariate development

A pool of covariates was developed that were intended to represent relief, organisms, parent material, and neighbourhood according to the *scorpan* method (McBratney et al., 2003). The covariates developed were altitude, convexity, plan curvature, profile curvature, negative openness, SAGA wetness index (SWI), slope gradient, slope length factor (LS factor), stream power index (SPI), topographic position index (TPI), vertical distance to stream (VDS), soil adjusted vegetative index (SAVI; Huete, 1988), lithology, soil colour index (SCI), and soil redness index (SRI; Ray et al., *2004*). These covariates were selected from experience of soil modelling in the area and it was thought that they capture the environmental factors that affect soil depth sufficiently. The vegetative and soil indices were developed from the Sentinel-2A satellite obtained at a 10 m resolution. The SAVI was calculated from a September 23, 2016 image to capture the vegetative growth before harvest. The soil indices (SCI and SRI) were calculated from a February 03, 2017 image during the fallow period. Together with the lithology map, the soil indices were used as an indication of soil parent material. The bands used to calculate the indices, the equations to calculate the indices, and a description of the indices used are shown in Table 2. For the SAVI, the vegetative factor (L) was set to 0.1 as this is suitable for most agricultural fields (Rondeaux et al., 1996).

2.3.4. Soil depth class predictions

Multinomial logistics regression with L_2 regularised (ridge regression) maximum log-likelihood was the algorithm applied in DSMART; however, it cannot be assumed that an algorithm will work best for a particular dataset (Kuhn and Johnson, 2013). Therefore, DSMART was run with the original C5.0 algorithm (Quinlan, 1993) and the Random Forest (RF) algorithm (Breiman, 2001) as implemented in previous studies. Additionally, as soil depth classes are ordinal in nature, an ordinal logistics regression (OLR) was also run. However, MLR outperformed these algorithms for all performance measures (see Section 3.5).

Table 2
Sentinel-2A satellite bands obtained, the equations to calculate the indices, and a description of the indices.

Bands	Central wavelength (μm)	Resolution (m)
Blue (B)	0.490	10
Green (B)	0.560	10
Red (R)	0.665	10
Near infrared (NIR)	0.842	10

Indices	Equation	Property
Colour Index	$(R - G)/(R + G)$	Soil colour
Redness Index	$R^2/(B * G^3)$	Hematite
SAVI	$\frac{NIR - R}{NIR + R + L}(1 + L)$	Chlorophyll reflectance

As implemented in the glmnet R package (Friedman et al., 2010), MLR finds the minimum of the multinomial loss function through coordinate descent. Coordinate descent is a optimisation procedure whereby, the loss function is minimised in a step wise processes corresponding to the coordinates of a hyperplane (Wright, 2015). Ridge regression shrinks covariate coefficients towards each other when covariates are correlated by panelising the coefficients. This is as opposed to LASSO which shrinks covariate coefficients towards zero and performs an imbedded feature selection. The degree of penalty is controlled by the λ value. Essentially, ridge regression slightly increases the bias of the model to decrease the variance, minimises the effect of collinearity, and prevents overfitting (Kuhn and Johnson, 2013). Due to these properties, ridge regression was selected and the λ value was optimised for each realisation. A ridge regression was chosen over LASSO because it is computationally faster and easier to implement.

Besides the algorithm chosen and sampling method, this implementation of DSMART follows the process described by Odgers et al. (2014). DSMART was run with 100 realisations and a minimum of 15 random samples per polygon. The random samples were based on weights determined by the depth class probabilities specified in the LTS legend. Area proportional sampling was conducted so polygons with a larger area, also have a larger number of random samples. However, because both the LTS-GM5 and the LTS-EX5 have 12 polygons, and the LTU only has three, the LTU was run with a minimum of 60 samples per polygon. This was so each polygon had a similar number of resamples for each realisation. Soil depth class probabilities were calculated from counting the number of times each pixel was classified as a particular soil depth class, for all realisations. This method is similar to the method performed by Kempen et al. (2000), except the MLR was implemented through a ridge regression and the original soil legacy data was spatial scaled through LFEs before predicting soil attributes.

2.3.5. Validation procedure

There were two sampling schemes developed to cover terrain attributes as best as possible. The two sampling schemes were expert samples and conditioned Latin hypercube sampling (cLHS; Minasny and McBratney, 2006). The location of both sample designs is shown in Fig. 5. In total, 93 soil observations were classified and sampled of which, 47 were expert samples and 46 were cLHS. Expert samples were placed by an expert through a google satellite image overlaid with 5 m contour lines. The expert samples were predominantly placed on altitude transects. Conditioned Latin hypercube sampling was implemented through the clhs R package (Roudier, 2011). The cLHS was performed with 100,000 iterations on six different environmental covariates. The covariates selected were aspect, slope, plan curvature, profile curvature, SWI, and the SAVI.

To evaluate model performance, the depth to any root limiting layer was measured for all 93 soil observations and these observations were used to calculate the kappa coefficient, overall accuracy, producer accuracy (PA), and user accuracy (UA) of the first most probable class raster. This is a deterministic accuracy assessment of the model performance. Additionally, the models were evaluated on the combined accuracies of the two most probable class rasters. This can be seen as a stochastic accuracy assessment of model performance.

To evaluate the spatial uncertainties, a confusion index (CI) between the first and second most probable soil depth class rasters was created. This follows the CI developed by Burrough et al. (1997). The equation for the CI is shown in Eq. (2), where P_{max} is the probability of

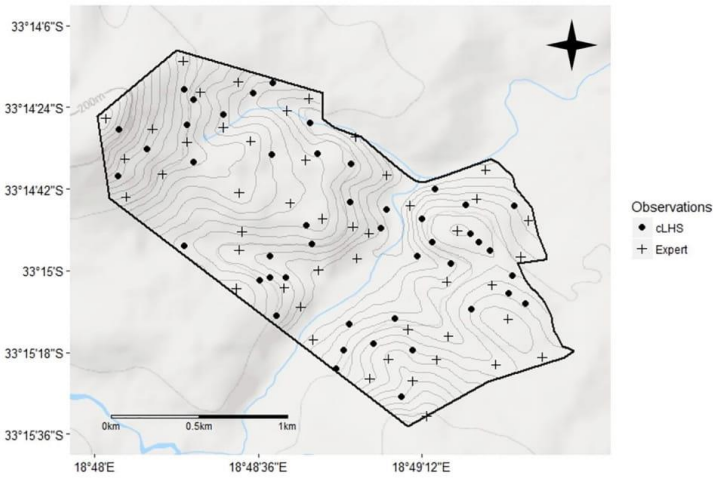


Fig. 5. Research site with expert placed and cLHS sample locations shown within 5 m contour intervals.

the most probable soil depth class and P_{max-1} is the probability of the second most probable soil depth class. This creates a CI raster where the closer a pixel is to zero, the more certain the model is of having correctly classified the soil depth class at that pixel. In other words, the larger the difference between the first and second most probable soil depth class probabilities, the more certain DSMART is of the predictions made.

$$CI = 1 - (P_{max} - P_{max-1}) \quad (2)$$

3. Results and discussion

3.1. Depth class probabilities

The probabilities of each soil depth class, on each LFE, according to the LTS legend can be seen in Table 3. The soil depth class probabilities align with the "mental model" of soil scientists. For example, the land types show a decreasing trend for shallow soils from crest to valley positions and deep soils show an increasing trend downslope. Moderately deep soils do not have a clear trend for land types Ac30 and Fb97 and show an increasing trend from crest to valley for land type Fb96. These probabilities are important for the final predictions as if the LTS legend is inaccurate, the final predictions will have a larger error. This was seen by Holmes et al. (2015) who stated that poorly delineated soil polygons extensively affect the outcome of the final predictions. This potential limitation can be overcome by changing the probabilities specified in the LTS legends. Vincent et al. (2018) successfully applied expert rules into the DSMART structure. However, this will require an expert pedologist familiar with a given site and places where financial resources are low, it is unlikely to find such an expert. Therefore, the probabilities were not changed as to evaluate the raw data specified in the LTS.

Table 3
Effective rooting depth probabilities for each terrain unit according to the LTS.

	Crest (%)	Mid-slope (%)	Foot-slope (%)	Valley (%)
Ac30				
Deep	0	5	40	50
Moderate	45	65	40	45
Shallow	55	30	20	5
Fb96				
Deep	0	5	10	38
Moderate	21	32	46	47
Shallow	79	63	44	15
Fb97				
Deep	0	7	15	95
Moderate	20	43	48	0
Shallow	80	50	37	5

3.2. Predicted landform elements

The proportion of LFEs predicted by the LTS-GM5 and LTS-EX5 relative to the LTS legend is shown in Fig. 6. The LTS-GM5 underestimated crest positions by 23% while the LTS-EX5 represented this landscape position well. Mid-slope positions had the highest proportion of area for all predictions. This was expected as sloping positions are in general, the most widely seen LFE (Libohova et al., 2016; Raska, 2012). However, the LTS-GM5 overestimated mid-slope by 30% and the LTS-EX5 underestimated these positions by 47%. The LTS-GM5 underestimated foot-slope positions by 53%, and the LTS-EX5 overestimated these positions by 34%. The LTS-GM5 predicted valley positions well and the LTS-EX5 overestimated this position by 64%.

3.3. Covariates and covariate importance

The covariates selected to establish soil environmental relationships and summary statistics of the covariate importance for all 100 realisations of the MLR LTS-GM5 model are shown in Table 4. Covariate importance is defined as the scaled (0–100%) absolute value of the coefficients and is the average over all realisations. The covariates were selected by running DSMART and eliminating the covariates which had the lowest importance until an acceptable accuracy was achieved or the accuracy started to decrease substantially. This was a time-consuming process which was thought to be improved by spatial principle component analysis. However, when running spatial principle component analysis to represent 95% of the covariate variability for all covariates, the model accuracy decreased substantially.

On average, convexity was the most important covariate followed by slope gradient, TPI, lithology, altitude, LS factor, and convergence index. Shallow and deep soils correspond to similar covariates, but for different reasons. Positions with low values for convexity and high values for topographic position and slope gradient correlate best with

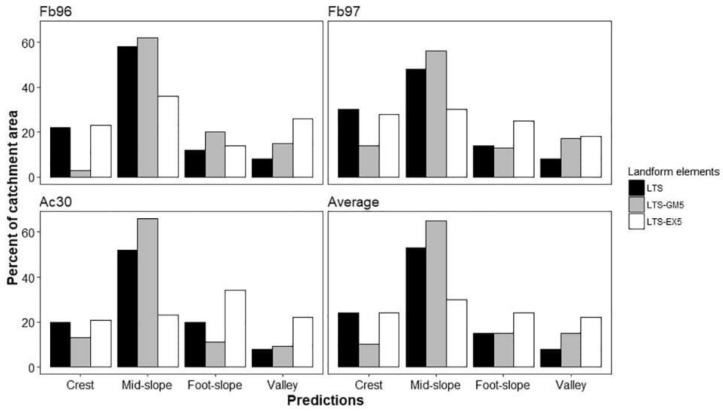


Fig. 6. Landform element proportion of area for the LTS-GM5 and LTS-EX5 compared to the LTS.

Table 4
The MLR model mean covariate importance for all realisations according to each depth class.

Covariates	Mean (%)	Shallow (%)	Moderate (%)	Deep (%)
Convexity	64	74	22	97
Gradient	41	40	24	60
TPI	39	48	14	54
Lithology	27	19	36	28
Altitude	25	29	19	27
LS factor	23	26	18	26
Convergence index	22	23	17	26

the processes controlling shallow soils. Deep soils correlate to processes controlling deposition and sedimentation, such as high convexity values and low slope gradient and TPI. Moderately deep soils correspond best to lithology; however, in general, covariates correlate less to moderately deep soils. This can be explained by moderately deep soils representing the transition from shallow to deep. This makes it more difficult for the algorithm to find covariates which separate moderately deep soils from the other soil (Burrough et al., 1997). Therefore, the covariates did not separate the classes well enough and the soil depth classes correlated to the same covariates (Chaney et al., 2016). However, when running DSMART with additional covariates, the model accuracies did not improve.

One option to improve the separation of moderately deep soils would be to run a supervised feature selection on a larger pool of covariates; however, this will increase the computation time of the algorithm and without soil point data, a supervised feature selection can be un-reliable as the random samples might not align with the feature space sufficiently. Alternatively, soil depth class criteria can be derived through a data driven technique such as k-means clustering (Forgy, 1965). This can be seen as a pre-processing technique to find structure in the LTS legend and has been shown to improve prediction accuracy (Trivedi et al., 2015). However, this data set is too small for such clustering as Dolnicar et al. (2014) states that 70 observations are needed for sufficient cluster separation. The LTS legends for this site, only have 56 depth ranges specified.

3.4. Most probable class rasters

The first and second most probable class rasters produced by the LTS-GM5 are shown in Fig. 7 and the probabilities of each soil depth class are shown in Fig. 8. The first most probable class raster mimics what was described in the LTS legend. Shallow soils had the highest probability on crest positions which is a result of erosional processes exceeding depositional processes (Scholms et al., 1983). Mid-slope positions had both shallow and moderately deep soils as deposition starts to increase downslope. Foot-slopes and valleys had both moderately deep soils and deep soils. Depositional processes such as colluvial wash or aeolian deposits have covered the gravelly creep and highly weathered shale, thus increasing the soil depth downslope (Lambrechts, 1983).

The first most probable class raster did not predict the shallow soils found in the middle of valleys where Typic endoaquepts and Typic endoaqualfs are found. Although these soils are physically deep, they have a permanent water table and therefore have a shallow depth class. An explanation for this misclassification is that shallow soils in valleys were not specified in the LTS legend. When evaluating the two most probable class rasters together, these shallow soils were still misclassified. However, these soils were classified as moderately deep soils as opposed to deep soils. This is a lesser loss of information because moderately deep soils are closer in taxonomical space to shallow soils, than deep soils are. This is an example of how the stochastic approach of DSMART can be beneficial, especially when evaluating the spatial distribution of soils which vary in a small spatial scale.

3.5. Accuracy assessment

The kappa coefficient and percent accuracy of the first probable class rasters, and the combined accuracy of the two most probable class rasters are shown in Table 5. What is clear, is classifying LFEs (regardless of LFE classification) before running DSMART is required to achieve accurate results. This can be seen in the combined accuracies for all resample polygon models. Multinomial logistics regression is the top performing algorithm in terms of kappa coefficient and overall accuracy for the LTS-GM5 and LTS-EX5. This is opposed to many digital soil mapping studies which report many tree-based models achieving a higher accuracy than linear models. At this geographic location, spatial and feature scale, this does not appear to be the case as none of the algorithms outperform another when taking the overall accuracies from all 100 realisations according to a student *t*-test ($p < 0.05$). Therefore, it is recommended that the simplest algorithm be optimised to increase computational efficiency. Due to the low computation demand, consistently high kappa coefficient and overall accuracy of the first most probable class rasters produced from the LTS-GM5 and LTS-EX5 models, the results produced from the MLR algorithm will be discussed further.

The LTS-GM5 and LTS-EX5 models had similar and satisfactory

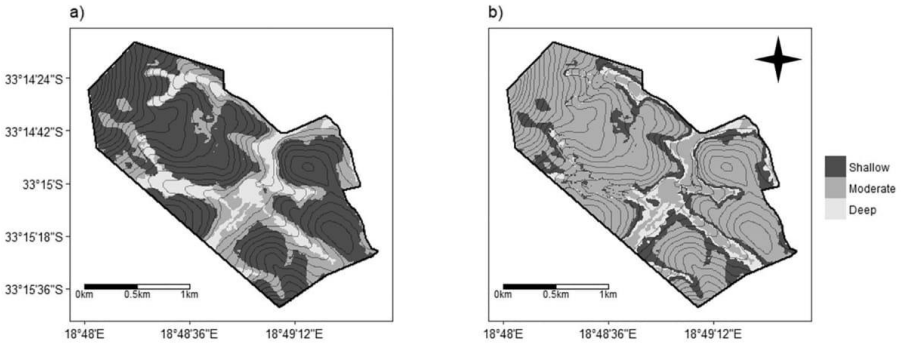


Fig. 7. First most probable raster (a) and second most probable raster (b) of the LTS-GM5 model shown with 5 m contour intervals.

results which greatly outperform the LTU model. Both models' first most probable rasters had a kappa coefficient indicating fair agreement according to Landis and Koch (1977) and accuracies similar or greater than, traditional soil map accuracies of 65% described by Marsman and de Gruiter (1986). When combining the accuracies of the two most probable class rasters, the LTS-GM5 and the LTS-EX5 models were comparable to the acceptable soil map accuracy of 85% according to Jensen (1986).

The satisfactory results were not expected for two reasons. First, it was assumed that the accuracies would decrease below an acceptable level without using soil point data to train the model, as previously found by van Zijl et al. (2013). Secondly, the LTS was developed on a regional scale, and this study focused on the farm-scale. Therefore, the probabilities predicted from the LTS may not be represented at the farm-scale, and the variability of soil depth should increase when re-scaling (McBratney, 1998). Fortunately, this does not seem to be the case. One explanation for this could be the fact that the area is covered by not one, but three land types. It is thought that this had an averaging effect when predicting DSMART across the different land types.

The increased complexity of the LTS-GM5 and LTS-EX5 also contributed to these resample polygons outperforming the LTU. A reason for this is that the polygons generated by LFEs have more detailed spatial soil information than the original land types. It is believed that this increased complexity stratifies soil depth classes into more homogenous sub-regions. This effectively aligned the random samples with the feature space which determines soil environmental relationships (Holmes et al., 2015). When the resample polygons' complexity is low, such as the LTU model, the random samples did not align with the feature space causing less accurate predictions. This result confirms that the two-step approach was necessary when predicting soil depth classes at the farm-scale.

Table 5
The kappa coefficient, overall accuracy, and the combined accuracies of the first and second most probable class rasters for all algorithms achieved for the LTS-GM5, LTS-EX5, and the LTU resample polygons.

	LTS-GM5	LTS-EX5	LTU
MLR			
Kappa	0.39	0.39	−0.13
Accuracy (%)	68	68	53
Combined (%)	90	90	67
OLR			
Kappa	0.27	0.31	0.00
Accuracy (%)	61	64	63
Combined (%)	80	86	67
C5.0			
Kappa	0.21	0.27	0.11
Accuracy (%)	63	61	64
Combined (%)	87	91	67
RF			
Kappa	0.21	0.25	0.02
Accuracy (%)	59	60	57
Combined (%)	87	89	65

The confusion matrix and the PA and UA for the LTS-GM5 model's first most probable soil class raster is shown in Table 6. Shallow soils were classified with the highest accuracy followed by deep soils and moderately deep soils. These results are not surprising as shallow soils have the largest probabilities specified in the LTS legends. Deep soils

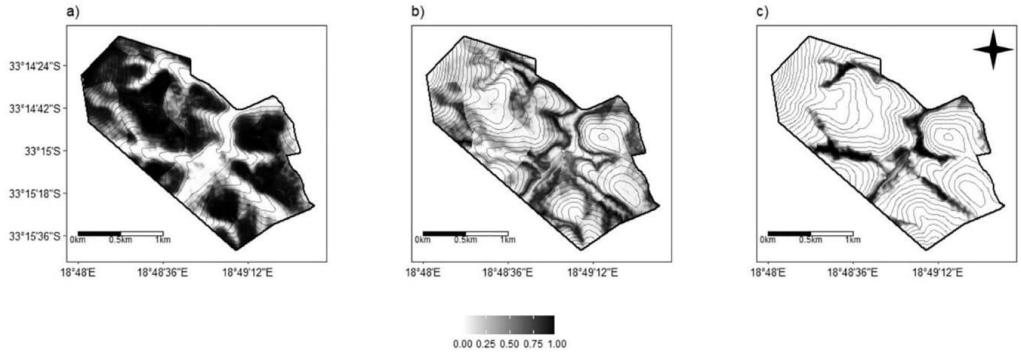


Fig. 8. Class probability rasters for shallow (a), moderately deep (b), and deep soils (c) shown with 5 m contour intervals.

Table 6
Confusion matrix with producer accuracy (PA) and user accuracy (UA) for the first most probable class raster based on the external evaluation from 93 soil profiles.

First	Deep	Moderate	Shallow	UA (%)
Deep	5	6	3	36
Moderate	3	10	7	50
Shallow	2	9	48	82
PA (%)	40	50	83	

were predicted with high accuracy; however, these soils were overly predicted and therefore, have a low UA. Moderately deep soils were classified with the least accuracy; however, moderately deep soils have a high UA. This indicates that moderately deep soils were under predicted but the predictions made, were classified with a high accuracy. Therefore, the deterministic accuracy assessment is limited to the probabilities and scale used in the LTS. To improve the predictions of soils with a lesser probability, expert rules could be assigned for each soil depth class to the resampling procedure (Odgers et al., 2014; Vincent et al., 2018).

When evaluating the combined accuracy of the two most probable class rasters, the LTS-GM5 model correctly classified 60% of deep, 100% of moderately deep, and 92% of shallow soils. Shallow and deep soil class accuracy improved by around 20% when assessing the accuracy of the combined rasters; however, moderately deep soil class accuracy improved by 50%. This can be attributed to the model overly predicting shallow and deep soils in the first most probable class raster and underestimating these soils in the second most probable class raster. Moderately deep soils showed the opposite trend leading to the large improvement when combining the accuracy of the two most probable class rasters. Therefore, DSMART's ability to produce multiple probability class rasters may improve results when downscaling to the farm-scale relative to deterministic approaches.

3.6. Spatial uncertainties

The CI for the LTS-GM5 is shown in Fig. 9. The CI is notable because it gives the uncertainties of soils in a spatial context which could assist in additional soil surveys and give insight into model performance (Odgers et al., 2014). The higher the CI, the more uncertain the model is of the soil depth class predicted at that given pixel. The LTS-GM5 had an average CI of 0.25 which is a slight improvement from the LTS-EX5 of 0.27 and a large improvement compared to the LTU of 0.57. The low

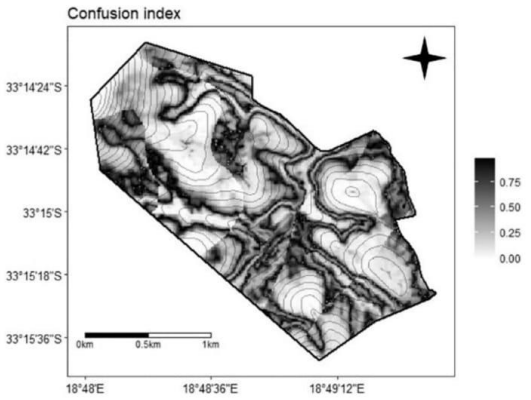


Fig. 9. Confusion index between the first and second most probable class rasters shown with 5 m contours.

average uncertainty suggests that the LTS-GM5 MLR model was the appropriate algorithm for this site. When running DSMART with OLR, C5.0 algorithm, and RF, the average CI increased to 0.32, 0.56, and 0.35 for the LTS-GM5, respectively.

It is evident that the uncertainties are spatially correlated and that there are patterns associated with soil depth class. Shallow soils had an average uncertainty of 0.15 and had the lowest uncertainty of any soil depth class. Moderately deep soils had the highest uncertainty of any soil depth class with an average of 0.64. Deep soils had an average uncertainty of 0.48. It was thought that the more accurate the model classified a soil depth class, the lower the uncertainties would be. This was the case when evaluating the first most probable raster, however, this was not the case when evaluating the two most probable rasters. For example, moderately deep soils had the highest uncertainty, however, when evaluating the two most probable rasters, moderately deep soils were correctly classified 100% of the time. An explanation for this is the weaker trend in the LTS data for moderately deep soils leading to the high uncertainties, although they were classified with high accuracy.

4. Conclusion

This study demonstrates a two-step disaggregation approach of a national resource inventory into a soil depth class map at the farm-scale. Landform elements were classified through the geomorphon algorithm and overlaid with the LTS. This created polygons with unique distribution of soil attributes. The polygons were used for resampling in DSMART implemented through a MLR to predict soil depth classes. This approach was compared with a manually delineated LFEs as resample polygons and disaggregating the LTS without LFE classification. The main findings of this study are:

- A soil depth class map was produced from a national resource map and achieved an accuracy similar or greater than, conventional soil survey accuracies.
- To disaggregate the LTS into farm-scale soil depth classes, initial LFE classification is required to produce polygons from which, DSMART can resample.
- Aggregated geomorphons can be used to predict LFEs specified in the LTS and therefore, can produce the first step and increase efficiency of the disaggregation process.
- Multinomial logistics regression implemented through DSMART, is capable of extracting soil depth classes from the probabilities specified in the LTS legend.
- The trend in the LTS legend greatly affects model performance and may be improved through incorporating expert knowledge of both LFEs and soils of the area.

This study has shown the potential to increase the accessibility and interpretability of the LTS. However, there is much work that needs to be done in terms of geomorphons and DSMART. This approach needs to be implemented in different geographic regions, on different scales, and with additional soil attributes. This may need to be addressed through standardising the implementation of this approach. A feature selection algorithm to select both geomorphons and covariates from the probabilities in the LTS may be an option. Then it has the potential to disaggregate the LTS across South Africa into soil maps that can be used for a variety of purposes.

Acknowledgements

This paper could have not been accomplished without the help of Dr. Freddie Ellis who developed the manually delineated landform elements. Also, his expert knowledge of soil formation and his contribution in developing the Land Type Survey in the area were paramount to this study. Additionally, we would like to thank the

contribution by Michael Flynn and Karen Angus for helping to edit the paper.

Funding

This research was funded by the National Research Foundation of South Africa reference SFH170525233469. NRF Building South Gate CSIR Complex, Meiring Naudé Road, Brummeria, Pretoria, South Africa. This research was also assisted by Stellenbosch University in the form of a travel grant. Private bagX1, Matieland 7602 Stellenbosch, South Africa.

References

Breiman, L., 2001. Random Forests. California, Berkeley.
Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103, 79–94.
Bunning, S., McDonagh, J., Rioux, J., 2011. Part 2 - field methodology and tools. In: Roam: FAO Manual for Local Level Assessment of Land Degradation, Sustainable Land Management and Livelihoods. 60.
Burrough, P.A., Van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77 (2–4), 115–135.
Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: a 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67.
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., et al., 2015. System for Automated Geoscientific Analysis (SAGA). Geoscientific Model Development.
Devia, G.K., Ganasri, B.P., Dwarakish, G.S., 2015. A review on hydrological models. Aquat. Procedia 4, 1001–1007.
Dolnicar, S., Grün, B., Leisch, F., Schmidt, K., 2014. Required sample sizes for data-driven market segmentation analyses in tourism. J. Travel Res. 53 (3), 296–306.
Dragut, L., 2011. Automated classification of topography from SRTM data using object-based image analysis. Geomorphometry 113–116.
Forgy, E., 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classification. Biometrics 21, 768–769.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–24.
GRASS Development Team, 2017. [Online], Available: http://grass.osgeo.org.
Grunwald, S., 2006. Environmental Soil-Landscape Modeling. Taylor & Francis, Boca Raton, Florida.
Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. Geoderma 185–186, 37–47.
Hengl, T., 2006. Finding the right pixel size. Comput. Geosci. 32, 1283–1298.
Hengl, T., Evans, I.S., 2009. Mathematical and digital models of the land surface. Dev. Soil Sci. 33, 31–63.
Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. CSIRO 53, 865–880.
Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25 (3), 295–309.
Irvin, B.J., Ventura, S.J., Slater, B.K., 1997. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. Geoderma 77 (2–4), 137–154.
Iwahashi, J., Pike, R.J., 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology 86, 409–440.
Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons — a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156.
Jensen, J.R., 1986. Introductory Digital Image Processing. Prentice Hall, Englewood Cliffs, NJ.
Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J., 2000. Updating the 1:50,000 Dutch Soil Map Using Legacy Soil Data: A Multinomial Logistic Regression Approach. vol. 97 Elsevier B.V.
Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands. Geoderma 241–242, 313–329.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York.
Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., et al., 2018. [Online], Available: https://cran.r-project.org/package=caret.
Lambrechts, J.J.N., 1983. In: Deacon, H.J., Hendey, Q.B., Lambrechts, J.J.N. (Eds.), Soils, Soil Process and Distribution in the Fynbos Region: An Introduction. Council for Scientific and Industrial Research, Pretoria.
Land Type Survey Staff, 1972-2002. Land Types of South Africa on 1:250 000 Scale. Pretori.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.
Libohova, Z., Winzele, H.E., Lee, B., Schoeneberger, P.J., Datta, J., Owens, P.R., 2016. Geomorphons: landform and property predictions in a glacial moraine in Indiana landscapes. Catena 142, 66–76.
MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets Syst. 113, 81–109.
Malone, B.P., Styc, Q., Minasny, B., McBratney, A.B., 2017. Digital soil mapping of soil carbon at the farm scale: a spatial downscaling approach in consideration of measured and uncertain data. Geoderma 290, 91–99.
Marsman, B.A., de Gruiter, J.J., 1986. Quality of Soil Maps, No. 15 ed. Netherlands Soil Survey Institute Soil Survey Papers, Wageningen, The Netherlands.
McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. Nutr. Cycl. Agroecosyst. 50, 51–62.
McBratney, A.B., Santos, M.L.M., Minasnya, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388.
Minasny, B., McBratney, A.B., 2016. Digital soil mapping: a brief history and some lessons. Geoderma 264 (August), 301–311.
Myburgh, P.A., Conradie, J.L., Zyl, V., Condradie, W.J., 1996. Effect of soil depth on growth and water consumption of young *Vitis vinifera* L. cv. pinot noir. S. Afr. J. Enol. Vitic. 17 (2), 53–62.
Nauman, T.W., Thompson, J.A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385–399.
Nauman, T.W., Thompson, J.A., Rasmussen, C., 2014. Semi-automated disaggregation of a conventional soil map using knowledge driven data mining and random forests in the Sonoran Desert, USA. Photogramm. Eng. Remote. Sens. 80 (4), 353–366.
Odgers, N., Malone, B.P., 2017. rdsmart: Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees (R Package Version 2.0.3).
Odgers, N.P., Sun, W., Mcbratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100.
Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, California.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. [Online], Available: www.R-project.org.
Raska, P., 2012. Biogeomorphologic approaches to a study of hillslope processes using non-destructive methods. In: Studies on Environmental and Applied Geomorphology. Elsevier Science B.V., Amesterdam, pp. 21–41.
Ray, S.S., Singh, J.P., Das, G., Panigrahy, S., 2004. Use of High Resolution Remote Sensing Data for Generating Site-specific Soil Management Plan. Vol. 34 [Online], Available: http://www.cartesianos.com/geodoc/isprs2004/comm7/papers/25.pdf.
Rondeaux, G., Steven, M., Baret, F., 1996. Optimization of soil-adjusted vegetation indices. Remote Sens. Environ. 55 (2), 95–107.
Roudier, P., 2011. clhs: A R Package for Conditioned Latin Hypercube Sampling.
Rozanov, A., Lessovaia, S., Louw, G., Polekhovsky, Y., de Clercq, W., 2017. Soil clay mineralogy as a key to understanding planation and formation of fluvial terraces in the South African Lowveld. Catena 156 (May), 375–382.
Sarmento, E.C., Giasson, E., Webster, E.J., Flores, C.A., Hasenack, H., 2017. Regional disaggregating conventional soil maps with limited descriptive data: a knowledge-based approach in Serra Gaúcha, Brazil. Geoderma Reg. 8, 12–23.
Schmidt, J., Hewitt, A., 2004. Fuzzy land element classification from DTMs based on geometry and terrain position. Geoderma 121, 243–256.
Scholms, B.H.A., Ellis, F., Lambrechts, J.J.N., 1983. In: Deacon, H.J., Hendey, Q.B., Lambrechts, J.J.N. (Eds.), Soils of the cape coastal platform. Council for Scientific and Industrial Research, Pretoria.
Scull, P., Franklin, J., Chadwick, O.a., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27 (2), 171–197.
Silva, S.H.G., Owens, P.R., Duarte de Menezes, M., Reis Santos, W.J., Curi, N., 2014. A technique for low cost soil mapping and validation using expert knowledge on a watershed in Minas Gerais, Brazil. Soil Sci. Soc. Am. J. 78 (4), 1310.
Silva, S., Menezes, M., Mello, C., Góes, H., Owens, P., Curi, N., 2016. Geomorphometric tool associated with soil types and properties spatial variability at watersheds under tropical conditions. Sci. Agric. 73 (4), 363–370.
Smith, S., Bulmer, C., Flager, E., Frank, G., Filatow, D., 2010. Digital soil mapping at multiple scales in British Columbia, Canada. In: Program and Abstracts, 4th Global Workshop on Digital Soil Mapping. 17.
Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334–345.
Trivedi, S., Pardos, Z.A., Heffernan, N.T., 2015. The Utility of Clustering in Prediction Tasks.
Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2018. Spatial disaggregation of complex soil map units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142.
Wiese, L., Ros, I., Rozanov, A., Boshoff, A., de Clercq, W., Seifert, T., 2016. An approach to soil carbon accounting and mapping using vertical distribution functions for known soil types. Geoderma 263, 264–273.
Wright, S.J., 2015. Coordinate descent algorithms. Math. Program. 151 (1), 3–34.
Yang, L., Jiao, Y., Fahmy, S., Zhu, A.X., Hann, S., Burt, J.E., 2011. Updating conventional soil maps through digital soil mapping. Soil Sci. Soc. Am. J. 75, 1044–1053.
Zhu, A.X., 1997. A similarity model for representig soil spatial information. Geoderma 77, 217–242.
van Zijl, G.M., Le Roux, P.A., Turner, D.P., 2013. Disaggregation of land types using terrain analysis, expert knowledge and GIS methods. S. Afr. J. Plant Soil 30 (2014), 123–129.

Appendix C

Geoderma 352 (2019) 171–180

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Comparing algorithms to disaggregate complex soil polygons in contrasting environments

Trevan Flynn[a], George van Zijl[b], Johan van Tol[c], Christina Botha[c], Andrei Rozanov[a], Benjamin Warr[d], Cathy Clarke[a,*]

[a] Department of Soil Science, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
[b] Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa
[c] University of the Free State, Nelson Mandela Avenue, 9301 Bloemfontein, South Africa
[d] BetterWorld Energy Ltd. Chitemwiko Close, Kabulonga, Lusaka, Zambia

ARTICLE INFO

Handling Editor: Alex McBratney

Keywords:
DSMART
Spatial disaggregation
Machine learning
Model comparison
National inventories

ABSTRACT

In South Africa, the only soil resource available with full spatial coverage is the national resource inventory. Disaggregating this polygon-based inventory, is thus a logical step to create more detailed soil maps covering the entire country. The polygons are large in area encompassing complex soil-terrain patterns and research into disaggregation techniques has been limited. This study aimed to compare 10 algorithms, implemented through a modified DSMART ("Disaggregating and Harmonizing Soil Map Units Through Resampled Classification Trees") model, in their ability to disaggregate two polygons into soil associations in two environmentally contrasting locations. One site had high relief and strong catenal sequences (eastern KwaZulu-Natal Province) and the other site had low relief and a strong geological control of soil types (northern Eastern Cape Province). The algorithms compared were based on previous studies which included k-nearest neighbour, nearest shrunken centroid, discriminatory analysis, multinomial logistics regression, linear and radial support vector machines, decision trees, stochastic gradient boosting, random forest, and neural networks. The method involves stratifying the polygons with landform elements, randomly sampling the landform elements, allocating the soil classes based on the resource inventory, and predicting soil associations across a stack of covariates. This was done in an iterative process, creating multiple realisations of the soil distribution. The performance of each algorithm was based on their kappa and uncertainties. It was found that in general, robust linear models which either utilise an embedded feature selection or regularise covariates, performed best. In the area with high relief and clear toposequences, nearest shrunken centroid was the top performing algorithm with a kappa of 0.42 and an average uncertainty of 0.22. In the area with relatively low relief and complex geology, the results were unsatisfactory. However, a regularised multinomial regression was the top performing algorithm, achieving a kappa of 0.17 and an average uncertainty of 0.84. The results of this study highlight the versatility of a technique to disaggregate South Africa's national resource inventory, where algorithms can be chosen on expert knowledge, model averaging can be performed, the top performing algorithm can be chosen, and algorithm parameters can be optimised.

1. Introduction

A large focus of digital soil mapping research in South Africa has been on disaggregating a single polygon in the Land Type Survey (LTS). The LTS covers the whole of South Africa and each polygon consists of an area with a relatively uniform climate, geology, topography and well-defined soil catena giving information on agriculture potential. However, the LTS consists of large soil-terrain polygons with very few georeferenced soil observations specified. These polygons were developed at a 1:250,000 scale, which is too coarse to inform land use management.

Disaggregation of soil data involves extracting the spatial location of each soil type from soil legacy data consisting of polygons with many soil types, thereby creating a more detailed soil map (Odgers et al., 2014; Thompson et al., 2010). This involves no or very few georeferenced soil observations to train models. There have been studies that compare several algorithms trained on point observations to predict soil type (Brungard et al., 2015; Heung et al., 2016). In contrast, there has

* Corresponding author.
E-mail address: cdowding@sun.ac.za (C. Clarke).

https://doi.org/10.1016/j.geoderma.2019.06.013
Received 22 February 2019; Received in revised form 3 June 2019; Accepted 9 June 2019

been little research comparing algorithms used to disaggregate national resource inventories in different environmentally contrasting areas. This is surprising as these inventories are seen as a wealth of information that often cover much, if not all, of a country.

Developed by Odgers et al. (2014), DSMART has become a popular disaggregation model because it predicts all soil types simultaneously and the predictions are not bound to the legacy polygons. Additionally, it is a stochastic approach which predicts the likelihood that a particular soil type is at a location, instead of concrete soil type boundaries. DSMART re-samples soil polygons, assigns the samples to a soil type based on the legacy data, and trains a decision tree on covariates. This creates multiple realisations of the soil distribution. The final product is a specified number of probability class rasters. It also produces spatial uncertainties (confusion) between the first and second most probable class rasters.

Decision tree based approaches have become popular in disaggregation models (Bui and Moran, 2001; Nauman and Thompson, 2014; Silva et al., 2016; Subburayalu et al., 2014). This is because decision tree algorithms can imitate the "mental model" of soil scientists and can handle either discrete or continuous data (Bui et al., 1999; Bui and Moran, 2001; Odgers et al., 2014). Decision trees can also handle non-linear relationships making them powerful classifiers (Breiman et al., 1984). One criticism of decision trees is that they are subject to overfitting (Grunwald, 2009) and therefore, other studies such as Häring et al. (2012), Nauman et al. (2014), Chaney et al. (2016), Vincent et al. (2016), and Møller et al. (2019) have used random forest to bypass such limitations (RF; Breiman, 2001). Random forest is also known for increasing model performance by reducing the variance (Buhlmann and Yu, 2002; Strobl et al., 2009). However, it was found that, when implemented through DSMART, a regularised multinomial logistics regression performed better than decision trees and RF when disaggregating LTS polygons at the farm-scale into a depth class map in South Africa (Flynn et al., 2019).

In the LTS legend, terrain information is specified as the percent area of landform elements on each land type. Specific landform elements found on each land type, are known as terrain morphological units (TMUs). Terrain morphological units were manually delineated from 1:50,000 topographic sheets and have five units. These TMUs include crest, scarp, mid-slope, foot-slope, and valley. Soil information is displayed as local soil series (local classification) and is the percent area on each TMU. All other soil information is associated with each soil series.

Soil information in the LTS is strongly tied to the influence each TMU exerts on soil formation. At some spatial scales and locations, this is logical as TMUs distinguish the boundaries between processes such as accumulation, deposition, and leaching potential (Evans, 2012). Therefore, both conceptual and DSM techniques have focused on utilising these relationships when disaggregating the LTS. However, this approach is problematic when topography is not the main driver of soil formation. For example, when land types cross many contrasting parent materials, parent material will exert a strong influence on soil formation as it affects both physical and chemical soil properties (Jenny, 1941).

Studies into disaggregating the LTS include Van Zijl et al. (2013), who disaggregated two land types in KwaZulu-Natal Province through an expert rules system and SoLIM software (Zhu, 1997). The authors concluded that a field survey was required to disaggregate the LTS as the disaggregation technique achieved a 35% accuracy. When adding observations to the model, the results drastically improved. The authors also concluded that adding lithology to the rules increased map usability but not map accuracy. Botha (2016), also used an expert knowledge approach whereby, TMUs were classified and assigned a dominant soil type based on the LTS data. The results were satisfactory for a high relief area (88%), but were un-satisfactory for an area controlled more by geology (37%). Flynn et al. (2019), disaggregated the LTS into soil depth classes at the farm-scale by stratifying LTS polygons with TMUs, and using the TMUs for re-sampling in DSMART. The study achieved a satisfactory accuracy of 68 to 90% depending on how many probability class rasters were used for evaluation.

Other studies that disaggregate complex soil-terrain polygons include Bui and Moran (2001), who used k-means clustering to classify soil associations and decision trees to classify fluvial facies which were strongly correlated to soil texture. The authors achieved an accuracy of 76 to 83% depending on the site in western New South Wales, Australia. Holmes et al. (2015), disaggregated soil-terrain polygons through DSMART in the whole of Western Australia. The authors implemented C4.5 decision trees and achieved an accuracy of 40% according to the three most probable class rasters and achieved a 71% accuracy when using higher levels of the soil classification system. However, these studies were conducted on much larger areas than the conventional approach in South Africa which focuses on disaggregating a single land type.

The aim of this study was to evaluate 10 algorithms on their ability to disaggregate the LTS in two environmentally contrasting areas (high relief, strong catenal sequence vs low relief, weak catenal sequence, and strong soil-geological relationships) using a modified DSMART model. The model allows the implementation of many classifiers and incorporates additional features that can be used to optimise the model for an area. This also has implications for disaggregating large datasets such as SOTER (Soil Terrain Dataset) (Dijshoon et al., 2008). Therefore, it has implications for further work in larger areas. This can be seen as an add-on to DSMART which can also be applied over different scales.

2. Method

2.1. Site description

Two land types were used in this study (Fig. 1), one site at Cathedral Peak in eastern KwaZulu-Natal Province (28° 30′S to 29° 30′S and 29° 00′E to 29° 30′E) and another site at Ntabelanga in northern Eastern Cape Province (31° 03′S to 29° 09′S and 28° 30′E to 28° 44′S). Both sites were selected due to their contrasting environments and data availability.

The Cathedral Peak site forms part of the South African National Environmental Network (SAEON) and consists of several protected, near pristine catchments. The site is located in the Drakensberg mountain range close to the border of Lesotho. It is the Ac265 land type (sheet 2828 Harrismith) encompassing 9.5 km^2 of relatively high relief and uniform geology comprising of basaltic rocks of the Drakensberg formation (Land Type Survey Staff, 1976–2002). Cathedral Peak has an Ustic climate with an average precipitation of 1130 mm (Schulze, 2007). The altitude of the study area, ranges from 1827 to 2068 m and is mainly covered by mesic grasslands interspersed with forest patches and wetlands.

The Ntabelanga site is the Db334 land type (sheet 3128 Umtata) encompassing 7.4 km^2 of relatively low relief ranging in altitude from 871 m to 1128 m with a complex geology. The geology consists of brownish-red and grey mudstone and sandstone of the Tarkastad Subgroup, Beaufort Group with dolerite intrusions (Land Type Survey Staff, 1976–2002). This area was part of the old Transkei Homeland. Agricultural production is classified as Maize Mixed Farming (Dixon et al., 2001) on state-owned land administered through the Tribal Authority system. Soils in the area are extremely susceptible to erosion, yet the site is earmarked for construction of the large multipurpose storage Ntabelanga dam in the Tsitsa River (Van Tol et al., 2014). The Ntabelanga area has a semi-arid climate with an average precipitation of 700 mm.

2.2. Legacy data

2.2.1. Land type terrain data

Each land type consists of unique patterns of terrain, soil, and

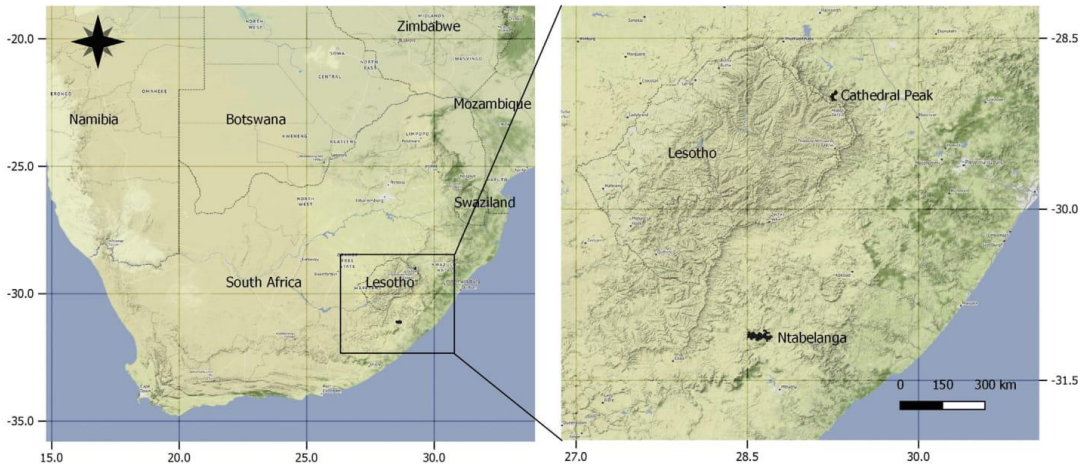


Fig. 1. The two study sites within southern Africa and zoomed into the eastern region of South Africa.

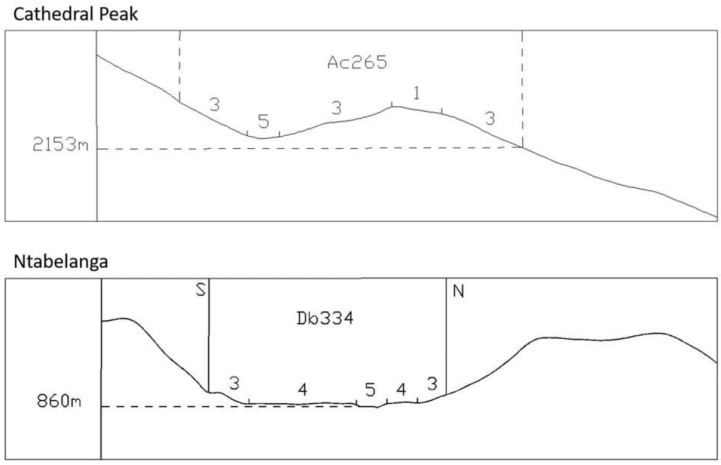


Fig. 2. TMUs situated on Cathedral Peak (Ac265) and Ntabelanga (Db334) taken from Land Type Survey Staff (1976–2002).

macro-climate (Paterson et al., 2015). A 2-dimensional depiction of the manually delineated TMUs on each land type are shown in Fig. 2. Cathedral Peak is dominated by mid-slope (3) consisting of 85% of the area, crest (1), and valley positions (5) make up 10% and 5% of the area, respectively. The slope at Cathedral Peak ranges from 4 to 60%. Ntabelanga is dominated by foot-slopes (4), encompassing 80% of the area, while mid-slope and valley positions consist of 10% of the area in total. The slope on the land type ranges from 4 to 10% (Land Type Survey Staff, 1976–2002).

2.2.2. *Land type soil data*

In the LTS, soil information includes soil type, depth, clay percent, textural class, and root limiting layer (Land Type Survey Staff, 1976-2002). Cathedral Peak has a clear toposequence consisting of a shallow association (Lithic Haplustepts and Lithic Humustepts), characterised by soils grading into bedrock within 50 cm from the soil surface on crest positions, the apedal association (Lithic and Typic Haplustox), comprised of generally deep soil with an apedal structure on mid-slopes, and a wet association (Typic Haplohumist), soils showing morphological signs of gleying in the valleys. Therefore, the soils were aggregated into these three soil associations based on lithic, oxic, and hydromorphic soil properties.

Ntabelanga's soils are more structured, complex, and are heavily controlled by geology/lithology. The soils were aggregated into apedal (Typic/Plinthic Haplustox), duplex (Typic Albaqualfs), semi-duplex or pedo (Typic Haplustalfs), shallow (Lithic Haplustept and Humustepts) and wet (Endo Aqualfs and Aquepts) associations. Soil associations were based on erosion potential. For example, apedal soils have a deep profile and high iron content making them resistant to erosion. In contrast, duplex soils have a discrete textural boundary due to their binary profile making them highly prone to crusting and erosion. The duplex and pedo soil associations differ in that the duplex soil association comprises soils with a prismatic structure in the subsoil, while the subsoil structure is angular blocky in the pedo soil association. As the names suggest, the shallow soil association grades into bedrock within 50 cm of the soil surface, while wet soil association shows

Table 1
Soil associations percent area on each TMU for Cathedral Peak and Ntabelanga.

Cathedral Peak			
Associations	Crest (%)	Mid-slope (%)	Valley (%)
Apedal	–	60	–
Shallow	100	40	–
Wet	–	–	100

Ntabelanga			
Association	Mid-slope (%)	Foot-slope (%)	Valley (%)
Apedal	10	15	–
Duplex	–	15	20
Pedo	30	40	40
Shallow	60	10	5
Wet	–	15	35

morphological evidence of water logging. The percent area of each soil association on the TMUs are shown in Table 1.

2.3. Model development

2.3.1. Polygon stratification

A topographic index landform classification (TPIc) was used to predict landform elements on each land type. The TPIc system compares the elevation at each pixel to the neighbourhood around that pixel (Weiss, 2000). For this study, the algorithm was implemented with a neighbourhood of 100 m for both land types. The landform elements were aggregated into TMUs according to Table 2. The TMU development is in contrast to the study by Flynn et al. (2019), whom introduced the use of geomorphons to segregate the landscape instead of TPIc. Additionally, in contrast to the previous work, this study focused on a regional scale. The aggregation was a subjective procedure based on the LTS specifications. For example, there is no foot-slope TMU at Cathedral Peak and no crest TMU classified on the Ntabelanga land type. Therefore, each land type was stratified into three polygons. This process is similar to that of Libohova et al. (2016), who aggregated a 10 unit pattern recognition algorithm known as geomorphons (Jasiewicz and Stepinski, 2013), into a five unit system based on slope.

2.3.2. Model training

The algorithms were trained and predicted using a modified DSMART model in R software (R Core Team, 2017). The modified method incorporates the caret R package where different classification algorithms can be used (Kuhn et al., 2018). The caret R package also allows for optimisation such as cross-validation, different sampling techniques such as up-sampling, and pre-processing such as centring and scaling of the covariates. The R software code developed, can be found on GitHub. Besides the incorporation of the caret R package, the model is similar to that of the rdsmart package (Odgers and Malone, 2017). For n realisations, the modified DSMART model is as follows:

1. Stratify the LTS with TMUs and prepare covariates.
2. Draw m random samples from each TMU.
3. Assign samples a soil type based on probabilities specified in the LTS.
4. Train an algorithm on covariates (caret R package).
5. Predict soil type across covariates.

Once n realisations have been trained and predicted:

1. Count number of times each pixel is classified a soil type.
2. Calculate probabilities based on counts (counts/total).
3. Determine soil type at each pixel

In this study, the method follows that of Flynn et al. (2019), 15 random samples were drawn for each TMU, assigned a soil association according to the LTS, covariate values were extracted (Section 2.3.3), and soil associations were predicted for 100 realisations. However, the sampling differs from that of Flynn et al. (2019), as the soil associations with the least probability (in the LTS) were up-sampled after drawing the 15 random samples. This was to account for class imbalances on each TMU.

The soil class assignment during the re-sampling procedure can be seen as a target-based approach on landscape rules (Odgers et al., 2014). This target-based approach is in contrast to other methods such as Häring et al. (2012) and Vincent et al. (2016), who used landscape rules to assign samples a soil type. Additionally, this approach is different than that of Møller et al. (2019), as it only uses landscape rules found in the original resource inventory.

Model development was an iterative process using 10 different algorithms shown in Table 3. The algorithms were largely selected based on the studies by Brungard et al. (2015) and Heung et al. (2016), who compared similar algorithms in three semi-arid regions in western USA and British Columbia, Canada, respectively. For detail into each algorithm, see Hastie et al. (2009) and Kuhn & Johnson (2013).

2.3.3. Covariates

Topographic covariates at each site were developed from a 30 m Advanced Land Observation digital elevation model (DEM). The DEM was used because of its superior vertical accuracy to other freely available DEMs. The resolution of the DEM was used to define the predictions final resolution. The covariates used to train the models were altitude, aspect, catchment area and slope, convexity, downslope (DC) and upslope curvature (UC), plan curvature, profile curvature, local curvature (LC), LS factor, multiresolution valley bottom flatness

Table 2
Aggregation of TPIc landform elements into TMUs for Cathedral Peak and Ntabelanga.

TPIc	Cathedral Peak	Ntabelanga
9	Crest	Mid-slope
8	Crest	Mid-slope
7	Crest	Mid-slope
6	Mid-slope	Foot-slope
5	Mid-slope	Foot-slope
4	Mid-slope	Foot-slope
3	Mid-slope	Foot-slope
2	Mid-slope	Foot-slope
1	Valley	Foot-slope
0	Valley	Valley
−1	Valley	Valley

Table 3
Classification algorithms used to predict soil associations at Cathedral Peak and Ntabelanga.

Algorithm	Type
k-nearest neighbour (KNN)	Distance based learner
Nearest shrunken centroid (NSC)	Distance based learner
Linear discriminatory analysis (LDA)	Simple linear model
Multinomial ridge regression (MRR)	Generalised linear model (L_2 regularised)
C5.0 decision trees (C5)	Tree based learner
Random forest (RF)	Multiple decision trees grown in parallel
Stochastic gradient boosting (SGB)	Multiple decision trees grown in sequence
Linear support vector machines (SVL)	Linear boundary learner
Radial support vector machines (SVR)	Radial boundary learner
Multilayer perceptron (MLP)	Multiple hidden layer neural network

Table 4
Spectral covariates obtained and developed at Ntabelanga.

Bands	Band origin (μm)	Symbol
Blue	0.490	B
Green	0.560	G
Red	0.665	R
Near infrared (NIR)	0.842	NIR

Indices	Equation	Property
Brightness Index (BI)	$(R^2 + G^2 + B^2)/3^{0.5}$	Reflectance
Coloration Index (CI)	$(R - G)/(R + G)$	Soil colour
Redness Index (RI)	$R^2/(B * G^3)$	Hematite
Saturation Index (SI)	$(R - B)/(R + B)$	Spectral slope
NDVI	$(NIR - R)/(NIR + R)$	Chlorophyll

(MRVBF), negative openness (NO), SAGA wetness index (SWI), sky view factor, slope, terrain factor, and terrain roughness. All topographic covariates were developed in the System for Automated Geoscientific Analysis (Conrad et al., 2015). These covariates were thought to describe the topography of both land types sufficiently.

In addition to topographic covariates, spectral covariates were developed at Ntabelanga from the Sentinel 2A satellite and were mean aggregated into a 30 m resolution. The addition of spectral covariates was done with the knowledge that the soils are controlled less by topography. The spectral bands and indices can be seen in Table 4. These covariates were thought to represent soil, vegetation, and parent material according to the *scorpan* method (McBratney et al., 2003). A description of the spectral indices can be found in Bannari et al. (1995) and Ray et al. (2004).

2.4. Model evaluation

2.4.1. Field observations

Field observations at Cathedral Peak and Ntabelanga were conducted during previous studies detailed below and shown in Figs. 3 and 4, respectively. Soils in both surveys, were classified according to South African Soil Taxonomy (Soil Classification Working Group, 1991). It should be noted that the original LTS soil profiles were not considered in either land type as they were not specified in the LTS legend.

At Cathedral Peak, 58 stratified random samples were targeted which were stratified between the Cathedral Peak research catchments within the land type (Van Zijl and Botha, 2016). The profiles were classified by 50 expert participants from the South African Soil Surveyor's Organization (SASSO) working group. Even though the samples were clustered within the research catchments, they are deemed to sufficiently represent the land type for evaluation of the models, as the soil distribution should be similar throughout the land type.

Eighty-seven soil profiles were classified and sampled in the Ntabelanga area as part of three projects. The first was to characterise the erosion susceptibility of soils adjacent to the proposed Ntabelanga dam (Parwada and Van Tol, 2017), the second to determine the pollution from pit latrines to streams (Mamera and Van Tol, 2018), and the third to quantify carbon stocks within the Ntabelanga dam footprint (Van Tol et al., 2018). The soil observations were located in and around the proposed footprint. As with the Cathedral Peak samples, these samples are sufficient to evaluate the different disaggregation models, because they follow the catena specified in the LTS.

2.4.2. Evaluation

All soil observations were performed completely independent of the LTS. Additionally, the small scale of samples enhances the actual evaluation as it captures the variability efficiently and large datasets are known for increasing the accuracy of identifying the predominant soil type. The number of samples used for evaluation is regarded as sufficient, as it compares well to the number of samples used in other DSM projects in the area, such as 60 for land type disaggregation, 52 for an expert knowledge approach and 48 for a machine learning approach (Van Zijl, 2019).

Each model was evaluated on their kappa statistics of the first most probable class raster and their average confusion between the first and second most probable class rasters. The kappa gives an indication on the algorithms goodness of fit while confusion tells how certain the model is of its predictions. Therefore, these two indices were used as the main indicator of model performance. The confusion values were calculated according to Burrough et al. (1997). Kappa was evaluated only on the first most probable class raster as this is a simplified model often necessary for decision making. The equation for the confusion can be seen below, where P_{max} is the probability of the first most probable soil association and P_{max-1} is the probability of the second most probable

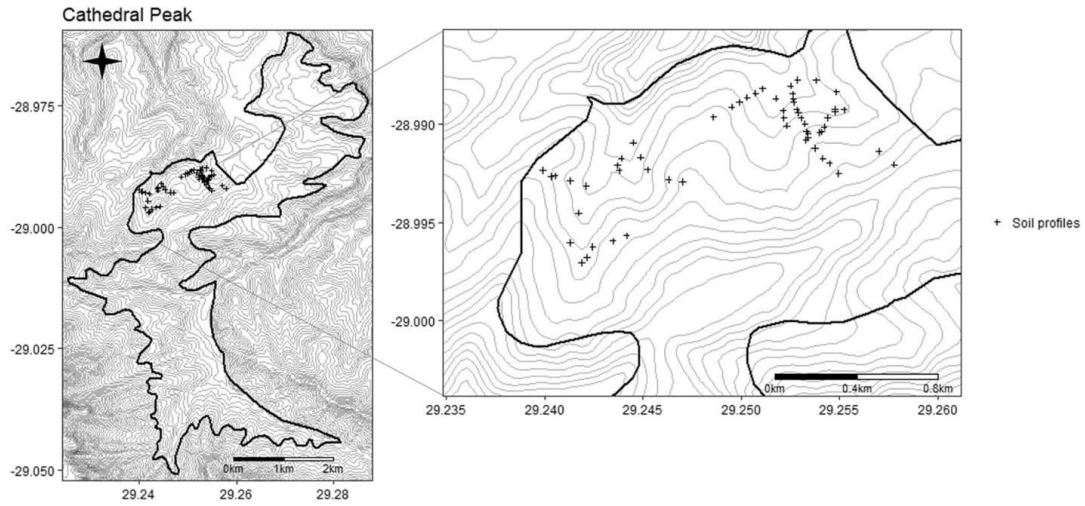


Fig. 3. Fifty-eight soil profiles in the Cathedral Peak land type (Ac265) shown on 20 m contour intervals.

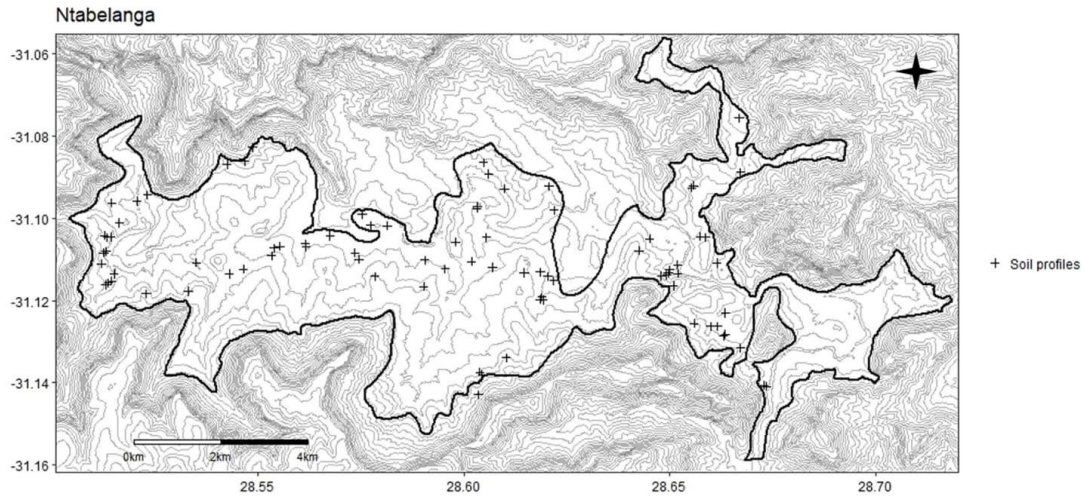


Fig. 4. Eighty-seven soil profiles in Ntabelanga land type (Db334) shown on 20 m contour intervals.

soil association at that pixel. The lower the confusion, the more certain the model is of the predictions. This measure is especially important where external soil observations are scarce, such as at Cathedral Peak.

$$Confusion = 1 - (P_{max} - P_{\max -1})$$

3. Results and discussion

3.1. Overall model performance

The kappa and confusion for each algorithm can be seen in Table 5. Over both land types, NSC had a competitive kappa and confusion. Nearest shrunken centroid has an embedded feature selection which minimises unimportant covariate centroids to zero (Tibshirani et al., 2003). Multinomial ridge regression had competitive kappa and confusion. However, the confusion was relatively higher at Cathedral Peak. The MRR algorithm panelises unimportant covariates based on the squared errors, thereby minimising the unimportant coefficients (Friedman et al., 2010). Therefore, these robust linear models prevent collinearity and overfitting which added to their performance.

Linear support vector machines had competitive kappa but had a comparatively higher confusion at Ntabelanga. Radial support vector machine performed well in terms of confusion, however, SVR had a relatively low kappa at Ntabelanga. Stochastic gradient boosting and RF also performed well in terms of confusion. This indicates that SVR, SGB, and RF although never the top performing algorithms, have little confusion between the first and second most probable soil associations across the land types. The original implementation through C5 performed well at Cathedral Peak in terms of kappa but performed poorly for all other performance indices.

Table 5
Algorithm performance showing kappa and confusion for Cathedral Peak and Ntabelanga.

Model	Cathedral Peak		Ntabelanga	
	Kappa	Confusion	Kappa	Confusion
C5	0.40	0.60	0.08	0.89
KNN	0.23	0.46	0.07	0.92
LDA	0.27	0.53	0.12	0.93
MLP	0.11	0.49	0.08	0.85
MRR	0.34	0.55	0.17	0.84
NSC	0.43	0.22	0.11	0.85
RF	0.26	0.45	0.07	0.85
SGB	0.42	0.49	0.09	0.83
SVL	0.41	0.54	0.11	0.90
SVR	0.36	0.35	0.05	0.83

Although most complex algorithms performed well, they may be unnecessary due to an already computationally heavy method, lack of improved results, and a decrease in interpretability. Additionally, algorithms such as SGB, MLP, and SVR failed to classify three out of the five soil associations at Ntabelanga. Random forest, C5, and SVR failed to classify two of the soil associations. This can be attributed to these algorithms over classifying the pedo association during each realisation. The over classification perpetuates through the realisations resulting in a great under representation of the other soil associations. This was observed after up-sampling the soil associations with the least percent area. However, without up-sampling, this trend was amplified, and performance dropped for all models. In contrast, linear models such NSC, MRR, LDA, and SVL only failed to classify the apedal association at Ntabelanga increasing their performance measures.

From these results, it is clear that some algorithms will perform better in either kappa or confusion values. Therefore, an algorithm should be chosen based on availability of observations to evaluate on. For example, if there are a few observations in an area, then a model that minimises the confusion might be more appropriate. This approach is more notable in areas with little knowledge of the soil distribution. If there are many observations, then the soil scientist can use expert knowledge to choose an algorithm based on specific needs or run many models and choose the best one. The latter is more suitable when disaggregating smaller areas such as a single land type. However, it was found that evaluating one realisation was a good indicator of model performance. Therefore, in the case of large areas, one realisation can be evaluated and then the best model can be selected.

3.2. Site specific model evaluation

3.2.1. Cathedral Peak

Nearest shrunken centroid achieved the highest kappa (moderate agreement), however SGB, SVL, and C5 had similar values. Where the NSC algorithm stands out is in the model confusion, where it is

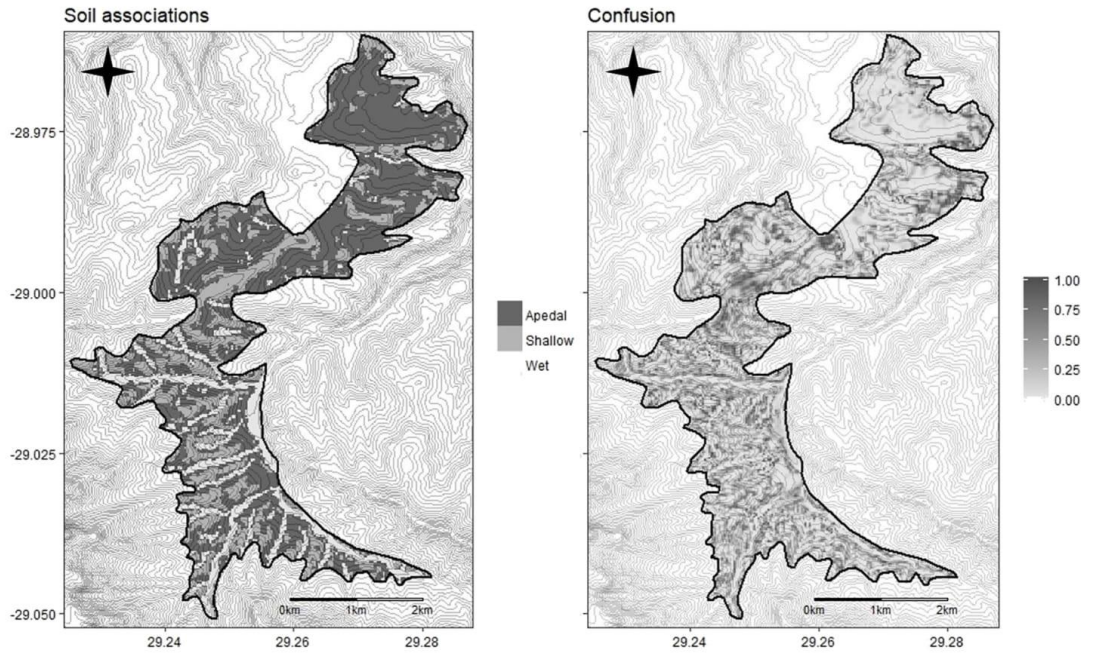


Fig. 5. Nearest shrunken centroid predictions and confusion at Cathedral Peak.

substantially lower than the next algorithm (SVR). Therefore, NSC is considered the best model for Cathedral Peak. The NSC predictions and confusion at Cathedral Peak are shown in Fig. 5. It should be noted that both C5 (76%) and KNN (73%) had a higher overall accuracy than NSC (71%), and SGB, SVL, and LDA achieved the same accuracy. Additionally, Botha (2016) achieved a kappa of 0.66 and an accuracy of 88% through an expert knowledge approach at Cathedral Peak. However, the approach presented in this study, is much more automated making it more suitable for larger areas (Van Zijl, 2019) and is not limited to the TMU boundaries (Odgers et al., 2014).

For NSC, the confusion is highly correlated to soil association and TMUs. To compare, the confusion was analysed by a post hoc Tukey-Kramer ($P < 0.05$*) implemented through a residual maximum likelihood model (REML). The REML model is necessary to account for spatial auto-correlation and is considered best practice with spatial data (Lark and Cullis, 2004). Wet soils had the lowest confusion of 0.18*, followed by apedal (0.21*), and shallow soils (0.25*). This is surprising, as wet soils were classified with less accuracy (50%) followed by that of apedal soils (68%). Shallow soils were classified with the highest accuracy of 85%which were found over both crest and mid-slope positions. Crests had the lowest confusion of 0.01* followed by valley (0.05*) and mid-slopes (0.25*). This indicates that soils found over many TMUs and TMUs with many soil associations will have the highest confusion.

At Cathedral Peak, most of the algorithms performed rather well. This was attributed to the clear toposequence and relatively non-complex soil pattern in the area. Shallow soils occur on crest positions where erosion exposes the lithic contact, apedal soils are found on mid-slopes where erosion is less, allowing more profile development, and wet soils occupy the valleys where water accumulates. These results largely confirm the LTS data and the study by Botha (2016), as soils are strongly tied to the TMUs on the land type.

3.2.2. Ntabelanga

There is no clear algorithm which substantially outperformed another at Ntabelanga. However, MRR achieved the highest kappa (slight agreement) but had similar confusion to SGB and SVR. The MRR model predictions and confusion can be seen in Fig. 6. The MRR model achieved an accuracy of 33% which is similar to that of Botha (2016), who reached a kappa of 0.20 and an accuracy of 37% at Ntabelanga. Additionally, these results are comparable to Van Zijl et al. (2013), who achieved an accuracy of 35% when disaggregating two land types with five soil associations. However, MRR overclassified both shallow and wet soils.

The MRR model's confusion had a similar trend to the NSC model at Cathedral Peak. The confusion was lowest for apedal soils of 0.56* followed by shallow (0.76*), wet (0.85*), pedo (0.92*) and duplex soils (0.95*). Soil associations which did not show a trend over the TMUs had the highest confusion. For example, shallow soils had a relatively low confusion but are found on every TMU, however, there is a clear trend from mid-slope to valley. In contrast, pedo soils are found on every TMU but with no clear trend. Therefore, pedo soils have a relatively high confusion. Additionally, all soil associations are found on foot-slopes which also had the greatest confusion (0.89*) followed by valleys (0.86*) and mid-slope positions (0.76*). This was expected and confirms TMUs with the greatest amount of soil associations will have the highest uncertainty.

In general, the models did not perform as well in Ntabelanga as in Cathedral Peak. Soil distribution patterns in the area are governed chiefly by the geology/lithology and secondary by the topography. Low model performance due to complex geological relationships was also found by Holmes et al. (2015) and is due to the TMUs not aligning the samples in the correct feature space. The horizontal and vertical variation within lithological layers e.g. sandstone, mudstone and dolerite within the Tarkastad and Beaufort groups, results in considerable variation in soils over short distances. This was clear during the field

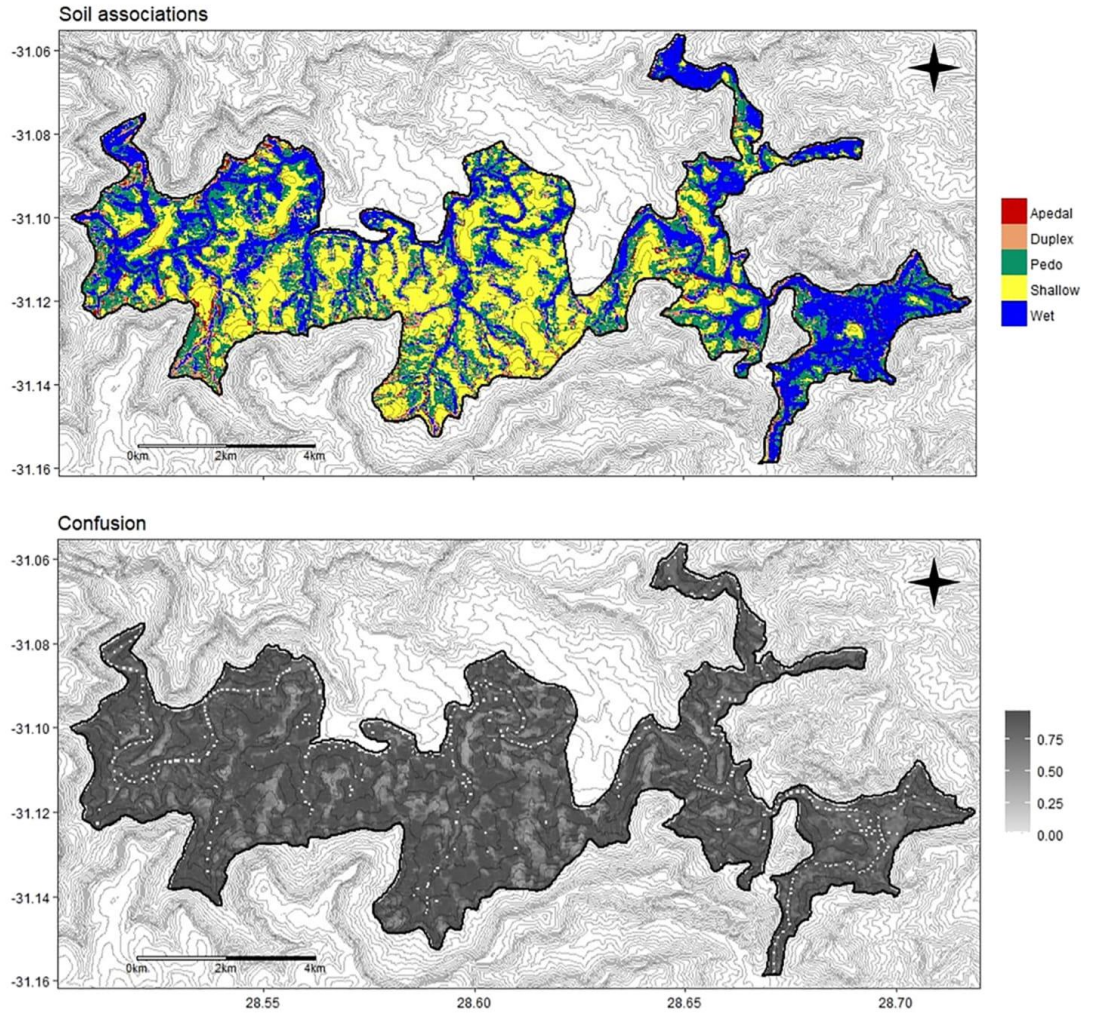


Fig. 6. Multinomial ridge regression predictions and confusion at Ntabelanga.

survey where it was observed that strongly structured soils (duplex and pedo) were found on mudstones and apedal soils were found on sandstones. Soils derived from dolerite, consisted of both red apedal and pedo soils. There is also colluvial material creating binary profiles which added to the complexity. In addition, the permeability of the different parent materials impact weathering patterns (depth) and soil/bedrock flow paths (wetness). This, together with the low relief, results in the occurrence of wet and shallow soils throughout the study area.

In an attempt to improve these results, model averaging of the overall top performing algorithms in terms of kappa (MRR, NSC, SVL, LDA, SGB) and model averaging using the algorithms which classified a particular soil association best (MRR, MLP, SGB, NSC), were tried. However, the results were disappointing only achieving a kappa of 0.05 and 0.06, respectively. Alternatively, if a geology/lithology map were available, expert knowledge could be used to determine the probability of each soil association on the different categories instead of using the LTS probabilities and TMUs. Additionally, expert rules could be used for soil type allocation during the resampling procedure as implemented by Häring et al. (2012) and Vincent et al. (2016). However, there was no reliable geology/lithology data available in an area to use as a covariate to train on. Although, the attempt failed, this shows the versatility of the modified DSMART approach.

Although results were poor at Ntabelanga, it shows some possibilities the modified DSMART model can perform. For example, running many algorithms and choosing the best one or using model averaging. It also provides additional functions that can be used to optimise algorithm parameters such as cross-validation at each realisation. Additionally, covariates can be pre-processed allowing for Box Cox and principle component analysis transformation. These additional techniques can be further explored when disaggregating a particular region.

Internationally, DSMART disaggregation produced accuracies comparable to other disaggregation studies. For example Møller et al. (2019), notes that 17–23% accuracy is the range when evaluating disaggregation approaches for the first most probable class raster when

Table 6
Five most important covariates and their descriptive statistics for the NSC and MRR algorithms at Cathedral Peak and Ntabelanga, respectively.

	Covariate	Mean (%)	Sd (%)	Lower (%)	Upper (%)
Cathedral Peak					
1	TPI	64	39	59	68
2	DC	62	32	59	66
3	LC	59	37	55	64
4	Convexity	43	26	40	46
5	SWI	39	23	37	42
Ntabelanga					
1	Aspect	36	27	34	38
2	Terrain factor	35	25	33	37
3	DEM	34	25	32	37
4	Convexity	31	23	29	33
5	NO	31	22	29	33

disaggregating over a large area with many soil types. Even with the insufficient results achieved at Ntabelanga, the accuracy was far above the international norm. This can be attributed to soil form aggregation and the scale of the site. Additionally, other research such as Møller et al. (2019), comes to the conclusion that a more detailed soil map is better than a less accurate soil map. When aggregating soils, it seems accuracy is more important than a detailed map in South African conditions. This is more important for land use management as the scale used, might not be appropriate for environmental modelling.

3.3. Covariate importance

The five most important covariates for NSC at Cathedral Peak and MRR at Ntabelanga are shown in Table 6. The covariate importance for NSC is calculated as the difference between a particular shrunken centroid for a class and that of the overall centroid (Tibshirani et al., 2003). Therefore, the larger the difference, the more important that covariate. The covariate importance for MRR is calculated on the absolute value of the coefficients. The further away from zero, the more important that covariate is. Both importance measures were averaged over all realisations and scaled to percentages.

At both sites, no covariate was the most important for all realisations. Therefore, the models are utilising different covariates for each realisation. This could indicate that the models' need many covariates to capture the soil distribution in the two areas. It was thought that this could also be a function of the number of samples per realisation. However, when performing the models with 50 samples per TMU, there was no difference in either model performance or covariate importance. Therefore, it could indicate that the covariates were collinear, and the number could have been reduced at Cathedral Peak, or that the covariates were not sufficient in the case of Ntabelanga.

Covariates which characterise slope position, slope shape, and water accumulation are the most important covariates at Cathedral Peak. This is no surprise as the soils are clearly controlled by landforms in the area. Apedal soils are correlated most with catchment slope as apedal soils are found on sloping positions. Both wet and shallow soils are strongly correlated with TPI as wet soils are in valley positions and shallow soils are on crest positions. However, the overall importance of catchment slope was low and TPI importance varied greatly with each realisation.

Covariates which characterise sun angle and amount, elevation, and slope shape correlate most with the soil associations at Ntabelanga. This is surprising as no spectral covariates were characterised even in the 10 most important covariates. It was thought that the spectral covariates would give more insight into the soil distribution, however, this was not clearly shown in these models. This was also seen by Møller et al. (2019), who found Landsat 8 bands and vegetative indices had a low importance. This could be due to the target-based soil assignment on TMUs which focuses more on topographic relationships. However, high soil erosion might have induced these results making the spectral covariates ineffective.

4. Conclusion

A modified DSMART model was developed to test 10 algorithms on their ability to disaggregate the LTS. The algorithms were compared on two environmentally contrasting land types in South Africa. The land types were first stratified with TMUs and these TMUs were used for resampling in DSMART. The algorithms were evaluated on the kappa of the first most probable class raster and the confusion between the first and second most probable class raster. The main findings of this study are:

- Robust linear algorithms such as NSC and MRR, were the top performing models for Cathedral Peak and Ntabelanga, respectively.
- When disaggregating a single land type, complex models do not improve the results and are less computationally efficient.
- Where there are strong soil-terrain relationships, the method produced satisfactory results such as at Cathedral Peak.
- Where there are strong soil-geological relationships, the method was deemed unfit such as in Ntablenga. Alternatively, another input map could be tried which does not focus on TMUS and relies more heavily on parent material.
- Grouping soil classes may be necessary when disaggregating soil maps with no legacy point data.
- Model averaging did not improve the results in the area with strong soil-geological relationships indicating the need to be supplemented with geological/lithological information.
- The results achieved, were comparable to other LTS disaggregation methods such as expert knowledge. However, this method is more automated making it more cost effective.

This study highlights the versatility the modified DSMART model brings to disaggregating the LTS. The modified DSMART allows users to choose the algorithm based on expert knowledge of an area, run many models to determine the best model, the ability to use model averaging, and/or optimise algorithm parameters. This methodology has implications for international datasets such as SOTER which also heavily relies on terrain to determine the soil distribution and which covers much of Southern Africa. This should be a priority in further research.

Acknowledgements

This work could not have been possible without the contribution of the Land Type Survey Staff (1976–2002), who developed the Land Type Survey. Specifically, A. L. Smith-Baillie and P. I. Steenekamp who developed the land types studied in this research. Additionally, we would like to thank B. Malone for publicly providing the DSMART code in R software which was modified for this study. We would also like to acknowledge The South African Soil Surveyor's Organization who allowed the Cathedral Peak dataset to be used in this research.

Funding

This research was funded by the National Research Foundation of South Africa, reference number SFH170525233469. NRF Building South Gate CSIR Complex, Meiring Naudé Road, Brummeria, Pretoria, South Africa. This study was also funded by South African National Environmental Network for the collaborated field samples at Cathedral Peak. Finally, the Water Research Commission for funding the soil observations at Ntabelanga.

References

Bannari, A., Morin, D., Bonn, F., Huete, A.R., 1995. A review of vegetation indices. Remote Sens. Rev. 13, 95–120. https://doi.org/10.1080/02757259509532298.

Botha, C.C., 2016. Disaggregating of Land Type Data to Aquire Functional Soil Information (MSc thesis). University of the Free State.
Breiman, L., 2001. Random Forests. Berkeley, California. https://doi.org/10.1017/CBO9781107415324.004.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth Int. Group, CA.
Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019.
Buhlmann, P., Yu, B., 2002. Analyzing bagging. Ann. Stat. 30, 927–961.
Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103, 79–94.
Bui, E.N., Loughhead, A., Corner, R., 1999. Extracting soil-landscape rules from previous soil surveys. Aust. J. Soil Res. 37, 495–508. https://doi.org/10.1071/S98047.
Burrough, P.A., Van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77, 115–135. https://doi.org/10.1016/S0016-7061(97)00018-9.
Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: a 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025.
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for automated geoscientific analysis (SAGA). In: Geoscientific Model Development, https://doi.org/10.5194/gmd-8-1991-2015.
Dijshoon, J., van Engelen, V., Huting, J., 2008. Soil and Landform Properties for LADA Partner Countries (Argentina, China, Cuba, Senegal, South Africa and Tunisia).
Dixon, J., Gulliver, A., Gibbon, D., 2001. Farming Systems and Poverty: Improving Farmers' LiveliHoods in a Changing World. FAO and World Bank, Rome and Washington DC.
Evans, I.S., 2012. Geomorphometry and landform mapping: what is a landform? Geomorphology 137, 94–106. https://doi.org/10.1016/j.geomorph.2010.09.029.
Flynn, T., Rozanov, A., Clercq, W. de, Warr, B., Clarke, C., 2019. Semi-automatic disaggregation of a national resource inventory into a farm-scale soil depth class map. Geoderma 337, 1136–1145. https://doi.org/10.1016/j.geoderma.2018.11.003.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–24. https://doi.org/10.18637/jss.v033.i01.
Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207. https://doi.org/10.1016/j.geoderma.2009.06.003.
Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. Geoderma 185–186, 37–47. https://doi.org/10.1016/j.geoderma.2012.04.001.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. Springer Series in Statistics.
Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014.
Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area Spatial Disaggregation of a Mosaic of Conventional Soil Maps: Evaluation Over Western Australia. vol. 53. CSIRO, pp. 865–880.
Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons - a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156. https://doi.org/10.1016/j.geomorph.2012.11.005.
Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, NY. https://doi.org/10.2307/211491.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New Yorkhttps://doi.org/10.1007/978-1-4614-6849-3.
Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, the R.C, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., 2018. Caret: Classification and Regression Training.
Land Type Survey Staff, 1976–2002. Land Types of South Africa on 1:250 000 Scale. (Pretoria, South Africa).
Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. Eur. J. Soil Sci. 55, 799–813. https://doi.org/10.1111/j.1365-2389.2004.00637.x.
Libohova, Z., Winzele, H.E., Lee, B., Schoeneberger, P.J., Datta, J., Owens, P.R., 2016. Geomorphons: landform and property predictions in a glacial moraine in Indiana landscapes. Catena 142, 66–76. https://doi.org/10.1016/j.catena.2016.01.002.
Mamera, M., Van Tol, J.J., 2018. Application of hydropedological information to conceptualize pollution migration from dry sanitation systems in the Ntabelanga Catchment Area, South Africa. Air Soil Water Res. 11, 1–12. https://doi.org/10.1177/1178622118795485.
McBratney, A.B., Santos, M.L.M., Minasnya, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.
Møller, A.B., Malone, B., Odgers, N.P., Beucher, A., Vangsø, B., Humlekrog, M., Minasny, B., 2019. Improved disaggregation of conventional soil maps. Geoderma 341, 148–160. https://doi.org/10.1016/j.geoderma.2019.01.038.
Nauman, T.W., Thompson, J.A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385–399. https://doi.org/10.1016/j.geoderma.2013.08.024.
Nauman, T.W., Thompson, J.A., Rasmussen, C., 2014. Semi-automated disaggregation of a conventional soil map using knowledge driven data mining and random forests in the Sonoran Desert, USA. Photogramm. Eng. Remote. Sens. 80, 353–366. https://doi.org/10.14358/PERS.80.4.353.
Odgers, N., Malone, B.P., 2017. Rdsmart: Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees (R Package Version 2.0.3).
Odgers, N.P., Sun, W., Mcbratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100. https://doi.org/10.1016/j.geoderma.2013.09.024.
Parwada, C., Van Tol, J., 2017. Soil properties influencing erodibility of soils in the Ntabelanga area, Eastern Cape Province, South Africa. Acta Agric. Scand. B Soil Plant Sci. 67, 67–76.
Paterson, G., Turner, D., Wiese, L., Van Zijl, G., Clarke, C., Van Tol, J., 2015. Spatial soil information in South Africa: situational analysis, limitations and challenges. S. Afr. J. Sci. 111, 1–7. https://doi.org/10.17159/sajs.2015/20140178.
R Core Team, 2017. R: A Language and Environment for Statistical Computing.
Ray, S.S., Singh, J.P., Das, G., Panigrahy, S., 2004. Use of high resolution remote sensing data for generating site-specific soil mangement plan. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.
Schulze, R.E., 2007. South African Atlas of Climatology and Agrohydrology, WRC Report 1489/1/06, Section 1.3. Water Research Commission, Pretoria, South Africa.
Silva, S.H.G., Menezes, M.D. de, Owens, P.R., Curi, N., 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma 267, 65–77. https://doi.org/10.1016/j.geoderma.2015.12.025.
Soil Classification Working Group, 1991. Soil Classificatioin: A Taxonomic System for South Africa, 2nd ed. Department of Agricultural Development, Pretoria, South Africa.
Strobl, C., Malley, J., Tutz, G., 2009. Introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–348. https://doi.org/10.1037/a0016973.
Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334–345. https://doi.org/10.1016/j.geoderma.2013.08.018.
Thompson, J.A., Prescott, T., Moore, A.C., Bell, J., Kautz, D., Hempel, F., Waltman, S.W., Perry, C.H., 2010. Regional Approach to Soil Property Mapping Using Legacy Data and Spatial Disaggregation Techniques, in: 19th World Congress of Soil Science. Soil Solutions for a Changing World. Brisbane, Australia, pp. 1–6.
Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2003. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Stat. Sci. 18, 104–117. https://doi.org/10.1163/15718093-12341267.
Van Tol, J.J., Akpan, W., Kanuka, G., Ngesi, S., Lange, D., 2014. Soil erosion and dam dividends: science facts and rural 'fiction' around the Ntabelanga dam, Eastern Cape, South Africa. South Afr. Geogr. J. 98, 169–181. https://doi.org/10.1080/03736245.2014.977814.
Van Tol, J.J., Akpan, W., Maroyi, A., Mutengwende, N., Huchermeyer, N., Ngesi, S., Nqandeka, H.M., Mamera, M., Bradley, G., Rowntree, K.M., 2018. The Mzimvubu Water Project: Baseline Indicators for Long-term Impact Monitoring. (WRC Proj. No. K5/2433).
Van Zijl, G., 2019. Digital soil mapping approaches to address real world problems in southern Africa. Geoderma 337, 1301–1308. https://doi.org/10.1016/j.geoderma.2018.07.052.
Van Zijl, G.M., Botha, J.O., 2016. In pursuit of a south African national soil database: potential and pitfalls of combining different soil data sets. South Afr. J. Plant Soil 1–8.
Van Zijl, G.M., Le Roux, P.A., Turner, D.P., 2013. Disaggregation of land types using terrain analysis, expert knowledge and GIS methods. South Afr. J. Plant Soil 30, 123–129. https://doi.org/10.1080/02571862.2013.806679.
Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2016. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142. https://doi.org/10.1016/j.geoderma.2016.06.006.
Weiss, A.D., 2000. Topographic position and landforms analysis. In: ESRI User Conference.
Zhu, A.X., 1997. A similarity model for representing soil spatial information. Geoderma 77, 217–242.