MODELLING SPATIAL VARIABILITY OF SOIL MOISTURE HOLDING CAPACITY IN A DRY SUB-HUMID LANDSCAPE

JACOB KAINGO

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY OF SOKOINE UNIVERSITY OF AGRICULTURE. MOROGORO, TANZANIA.

2017

EXTENDED ABSTRACT

Moisture scarcity is a limiting factor for sustainable agricultural productivity of dry sub-humid agroecosystemsof sub-Saharan Africa (SSA). Designing sustainable agricultural system management strategies responsive to the fluctuating soil moisture regime is essential. Detailed and accurate information on soil moisture storage conditions is essential for modelling agricultural system productivity. Moisture storage capacity of the soils is quantified by moisture holding capacity (MHC) which is defined as the difference between moisture content at field capacity (FC)and wilting point (WP). Data availability is limited for SSA due to high costs associated with direct measurement of MHC. Pedo-transfer functions (PTFs) and the digital soil mapping (DSM)framework offer an opportunity for characterising spatial variability of MHC through indirect approaches that integrate mathematical and statistical methods. Though various methods exist for prediction and mapping MHC, machine learning methods offer an avenue for more accurate characterisation of MHC. The main objective of this study was to improve understanding on estimation of soil moisture holding capacity at large spatial domains using machine learning algorithms. This was achieved througha probabilistic sampling scheme, development of MHC PTFs, and 3-dimensional characterisation of spatial variability of MHC. One hundred (100) sampling locations were established over a geographic area of about 44 km^{2}by k-means clustering using R-statistical software. Two sampling strategies were evaluated for optimisation of the sampling locations –a stratified random sampling (STRS) and spatial coverage sampling (SPCS). Bulk soil samples and soil cores were taken at three depth intervals of 0-30cm, 30-60 cm, and

60-100 cm at each sampling location. Geostatistical analysis and cross-validation were performed for assessment of the sampling schemes using root mean square error (RMSE), coefficient of determination (R^2) and Mean Error (ME) as indices. West-East anisotropy was evident in the MHC probably associated with topographic and land cover effects. Spatial dependence ratio for the stratified random sampling scheme (73 %) was higher than that of the spatial coverage sampling scheme (19 %). This implied that SPCSdesign had better spatial correlation than the STRS design due to a regular configuration of sampling nodes for SPCS design.Validation resultswere better for STRS design than SPCS design. Pedo-transfer functions were developed for FC and WP from support vector regression and multiple linear regression with soil physico-chemical properties as predictors. Support vector regression-PTFs had slightly better accuracy (RMSEs = 0.037 $cm^{-3}cm^{-3}$) than multiple linear regression PTFs (RMSEs = 0.038 $cm^{-3}cm^{-3}$) and other published PTFs. R^2 values for SVR-PTFs were 66.3 and 67.9 % while those for MLR-PTFs were 64.5 and 67.3% for FC and WP, respectively.Two machine learning algorithms (Random forests(RF) and cubist decision trees (CB)) combined with soil depth functions were evaluated for 3-dimensional mapping of MHC. Two DSM scenarios were also evaluated (Measured data only (DSM-A) and measured plusPTF-estimated data (DSM-B)).Principal component analysis was performed on spatial covariates layers representing soil forming factors for dimension reduction. Ten principal components with a cumulative variance > 70 % were selected for mapping process. Equal-area quadratic spline soil depth functions were fitted to model continuous vertical distribution of MHC data. Prediction accuracy was good with RMSEs ranging between 0.011-0.015 $cm^{-3}cm^{-3}$and R^2 between 36 - 81.4 %. Random forests had better accuracy than the Cubist decision trees. A RF-CB ensemble improves prediction accuracy.

DECLARATION

I, JACOB KAINGO, do herebydeclare to the Senate of Sokoine University of Agriculture that this dissertation is my own original work and that it has neither been submitted nor being concurrently submitted for degree award in any other institution.

JACOB KAINGO **DATE**

(PhD Candidate)

The above declaration is confirmed

PROF BONIFACE P. BILINYI **DATE**

PROF SIZA D. TUMBO **DATE**

COPYRIGHT

ACKNOWLEDGEMENTS

It has been a challenging endeavour! Not an endeavour accomplished entirely of my effort but of the selfless volition ofnumerous individuals who desiredto see me successfully finish this journey. Foremost, I extend praise and glory to God for keeping me in good healthand state of mind to accomplish the task.

My heartfelt gratitudeto my supervisors Prof Boniface P. Mbilinyi and Prof. Siza D. Tumbo for their endearingcontribution and patience for the successful accomplishment of this research work. They mobilised resources and secondedme for numerous opportunities to attend ascope of trainings to horn research skills. I have matured both professionally and socially with your core mentorship and guidance.It has been such a great privilege and humbling experience learning from you! I am also grateful to Dr Ludger Herrmann of the University of Hohenheim for his advice on the implementation of the fieldwork and the financial resources to support my research work.I thank Prof Dick J. Brus of Wageningen Univeristy for his guidance on designing my study and more so with the R-code. I am grateful to Prof Gerard Heuvelink for his insights and hosting me on my visit at World Soil Information – ISRIC (Wageningen, NL). Gratitude to Prof. Nganga I. Kihupi for reviewing and inputs for improvingsome of the draft manuscripts in this dissertation.

My doctoral studies were primarily sponsored by the Alliance for a Green Revolution in Africa (AGRA) through the Soil Health Training Programme financed by the Bill and Melinda Gates Foundation. I am very grateful to Dr Marie Rarieya who vouched for me to join the PhD program.I also acknowledgeDr Claudius Van de Vijver for the support accorded for the various Wageningen University-AGRA trainings. I benefited from additional financial support from the Trans-Sec project and Spurring a Transformation in Agriculture through Remote Sensing (STARS) Project. I am grateful to the coordinators of the Trans-Sec project Prof Frederick Kahimba, Dr Khamaldin Mutabazi of SUA and Dr Stefan Sieber, and Dr Frieder Graef of ZALF for the sponsorship. In the same vain, I also thank Dr Jan Dempewolf of University of Maryland who provided coordination for the STARS project. Warm gratitude to Pendo Meike Schaefer for organising my visit to the Institute of hydrology, ZALF and Dr Uwe Schindler for the insights into my work.I am grateful for the inspirationand encouragement from my PhD colleagues on the Trans-Sec project Hadijah Mbwana, Maureen Tatu Mnimbo, Yusto Yustas, Emmanuel Chilagane, Richard Festos, Paul Saidia, and Lutengano Mwinuka and office mates,Edmund Mutayoba, Isack Yona, and Dr Sixbert Mourice.

To my parents,James and Florence Kaingo, *mwanyala nabi!* for the numerous sacrifices for my education and upbringing. To all my siblings (Eva, Sarah, Grace and Chris), thank you for the fervent prayers and spurring me on to this goal.Special gratitude to my dear wife, Mariana L. N. Kaingo, for refuelling my spirits with the love, encouragement and thoughtful delicacies throughout this journey.

Not least, gratitude to small-scale farmers and my field assistants (Stanley Charles, Malik and John) in Ilakala village for their warmth, cooperation and toilfor me to successfully accomplish the fieldwork. And to all not listed here, thanks a mountain!

DEDICATION

To all those who have enabled me learn!

To my daughter Blessing Rehema Naula Kaingo.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND SYMBOLS

AACN	-	Altitude Above Channel Network
APSIM	-	Agricultural Production Systems sIMulator
ASE	-	Average Kriging Standard Error
ASP	-	Aspect
CB	-	Cubist algorithm
cm^3cm^{-3}	-	cubic centimetres per cubic centimetres
CV	-	Coefficient of variation
DEM	-	Digital Elevation Model
DLR-		*Deutsches Zentrum für Luft- und Raumfahrt e.V.* (German Aerospace Center)
DSM	-	Digital Soil Mapping
DSSAT	-	Decision Support System for Agrotechnology Transfer
FC	-	Field Capacity
GSIF	-	Global Soils Information Facility
LSF	-	Length Slope Factor
MAX	-	Maximum
ME	-	Mean Error
MHC	-	Mositure Holding Capacity
MIN	-	Minimum
MLA	-	Machine Learning Algorithm
MRBF	-	Multi-Resolution Valley Bottom Flatness Index
MRTF	-	Multi-Resolution Ridge-Top Flatness Index

PC	-	Principal Component
PCA	-	Principal Component Analysis
PLC	-	Planar Curvature
PRC	-	Profile Curvature
PTF	-	Pedo-Transfer Function
R	-	Correlation coefficient
R^2	-	Coeffiecient of determination
RF	-	Random Forest algorithm
RMSE	-	Root Mean Square Error
RMSSE	-	Root Mean Square Standard Error
SAGA	-	System for Automated Geoscientific Analyses
SDF	-	Soil Depth Functions
SLP	-	Slope
SPCS	-	Spatial coverage Sampling
SRTM	-	Shuttle Radar Topography Mission
STRS	-	Stratified Random Sampling
STWI	-	SAGATopographic Wetness Index
SVM	-	Support Vector Machines
SVR	-	Support Vector Regression
SW	-	Shapiro-Wilkinson Test Score
SWAT	-	Soil and Water Assessment Tool
TPI	-	Topographic Position Index
TWI	-	Topographic Wetness Index
UAV	-	Unmanned Aerial Vehicle

USGS	-	United States Geological Survey
WP	-	Wilting Point
WRB	-	World Reference Base
ZALF	-	Leibniz-Zentrum für Agrarlandschaftsforschung (ZALF) e. V.

CHAPTER ONE

1.0 GENERAL INTRODUCTION

1.1 Relevance of soil moisture holding capacity in drysub-humid regions

Dry sub-humids are a hotspot for food insecurity (Rockstrom *et al.,* 2007; Shiferaw *et al.,* 2014) due to their inherently complex ecological shifts characterised by climatic variability, low precipitation, and persistent moisture scarcity(Enfors *et al.,* 2008; Wang *et al.,* 2012).A substantial decrease in crop production of smallholder farms in dry sub-humid sub-Saharan Africa (SSA)is evident due to increased uncertainty of soil moisture supply,greatly contributing to crop loss and decline of livelihood quality (Shiferaw *et al.,* 2014; Reynolds *et al.,* 2015). Designing agricultural interventions to mitigate this existential agro-hydrologic challenge is essential for increased cropproduction to satisfy the rising food demand (Barron, 2004; Garrity *et al.,* 2012).Precise information on soil moisture holding capacity (MHC) is vital for understanding dynamics of the various interventions onproductivity of agricultural systems (Kilasara, 2010).

Soil moisture holding capacity (MHC) is the difference between moisture at field capacity (FC) and that at wilting point(WP) (Ladson *et al.,* 2006; Asgarzadeh*et al.,* 2010). MHC is soil moisture that is theoretically available for crop consumptive use (Ladson *et al.* 2006). Field capacity is the upper limit of MHC while WP is the lower limit. Measurements of moisture at FC and WP are conventionally done at predefined matric suctions. Wilting point is measured at a matric suction of 1500 kPa. However, there is no universally established matric suction for FC (Gaiser*et al.,* 2000; Asgarzadeh*et al.,* 2010). Matric suctions for FC reported in literature range from 5 kPa to100 kPa (Asgarzadeh*et al.*, 2010), mostly dependenton soiltexture (Asgarzadeh*et al.*, 2010) or geographical region (Gaiser*et al.*, 2000).For instance, the matric suction for FC is 5 kPa in the United Kingdom and 33 kPa in the United States. In tropical regions, 33 kPa has been used as standard matric suction for FC (Gaiser *et al.*, 2000). Moisture holding capacity is a necessary parameter for description of vadose zone fluid flow processes for water and nutrient management in agricultural systems (Santra *et al.*,2009; Vereecken *et al.*, 2010).

1.2 Spatial variability of MHC at large spatial domains

Mapping spatial variability of soil MHCat large scale is challenging (Levi *et al.*, 2016) as it exhibits a high degree of spatial variability at different scalesinfluenced by numerous factors (Si, 2008; Biswas *et al.*, 2013; Lark, 2016).Spatial variability of MHC considerably influences surface water storage, infiltration rate, evapotranspiration rate, and the recharge rates vary correspondingly, thereby creating uncertainties in assessment of the water balance for agricultural production systems (Zhu and Mohanty, 2006). Detailed and accurate characterisation of the spatial variability of MHC is thusinevitable for better agricultural water management (Xu *et al.*, 2009; Huang and Li, 2010).

Costs associated with direct determination of MHC at large scales like farming systems areprohibitive (Santra *et al.*, 2009; Rizzo *et al.*, 2016).Soil-landscape modelling is the most amenable approach at such large scale (Bou Kheir *et al.*, 2010; Odgers *et al.*, 2011). Conventional soil surveys have adopted a choropleth mapping approach that representssoil variability ashomogenousmutually exclusive contiguous polygons(Moore *et al.*, 1993; Odgers *et al.*, 2011; Adhikari *et al.*, 2013; Rizzo *et al.*, 2016).Soil variation is continuous and this approach of conceptual soil boundaries (Odgers *et al.*, 2015) does not sufficiently represent the intrinsic spatial variability of soil MHC (Moore *et al.*, 1993; Odgers *et al.*, 2011; Adhikari *et al.*, 2013; Taghizadeh-Mehrjardi *et al.*, 2015).

Digital soil mapping (DSM) is an efficient alternative for quantitative characterisation of spatial variability of MHC properties through the application of statistical and mathematical tools (McBratney *et al.*, 2003; Levi *et al.*, 2015). The advantage of the DSM approach is that the continuous spatial variability of MHC can be mappedwith a quantitatively defined soil-landscape model which does not heavily hinge on qualitative knowledge (Odgers *et al.*, 2011). DSM also has capacity to provide quantitative estimates of uncertainty of predictions(Odgers *et al.*, 2011; Odgers *et al.*, 2015). DSM can also integrate remote sensing products like DEMs for mapping soil MHC (Lagacherie *et al.*, 2013;Taghizadeh-Mehrjardi *et al.*, 2015).

1.3 Challenges and opportunities for DSM in smallholder farming systems

Implementing sustainable agricultural oragro-hydrological interventions for higher crop yields in smallholder cropping systems in SSA requires digital soil maps at increasingly finer scales (Lagacherie *et al.*, 2013; Hengl, *et al.*, 2017a).Unfortunately, MHC maps with a suitable scale for crop management in smallholder systems are scarce or often missing (Rizzo *et al.*, 2016; Nussbaum *et al.*, 2017). Although some DSM products have been developed for SSA (Leenars *et al.*, 2015; Hengl *et al.*, 2015; Hengl *et al.*, 2017a; Hengl *et al.*, 2017b), there is discordance of the spatial resolutions of these DSM productsfor resolving issuesat the farm scale (Malone *et al.*, 2017). Their grid cells are too coarse for meaningful on farm assessments (Malone *et al.*, 2017), ofsmallholder farm holdings of less than 1 hectare in Eastern Africa (Pender *et al.*, 1999).Accurate and fine-scaleDSM predictions of MHC propertieswould have better utility for decision support in smallholder systems(Nussbaum *et al.*, 2017; Hengl *et al.*, 2017a). For instance, to facilitate suitable crop enterprises selection adapted to the moisture limitations of their fields.

Accuracy ofDSM depends on availability of data (Odgers *et al.*, 2015) and distribution of sampling locations (Brus *et al.*, 2011). Data has,however, not been in adequate supply for SSA (Kilasara, 2010; Cambule *et al.*, 2014) with soil databases poorly developed for the region (Kilasara, 2010). Uniformly distributed and probabilistic sampling locations across the geographic space enhance precision and consistence of DSM predictions (Brus and Heuvelink, 2007; Walvoort*et al.*,2010). Pedo-Transfer Functions (PTFs) which use basicsoil physico-chemical data (e.g. texture, soil organic carbon) to estimate MHC (Zinn *et al.*, 2005;Vereecken *et al.*, 2010; Haghverdi *et al.*, 2012) have been integrated in a DSM framework to estimate missing MHC data (Leenars *et al.*, 2015). Some of the PTF techniques commonly used include multiple linear regressions (MLRs), support vector machines (SVMs), artificial neural networks (ANNs) (Khlosi *et al.*, 2016). However, existing PTFs are bound to unique development environments or have not been suitably conceived for application which results in propagation of uncertaintiesindigital maps (Chirico *et al.*, 2007).Remote sensing data is a cost-efficient means to overcome the lack of soil data that still severely limits DSM performance (Lagacherie *et al.*, 2013). Many

DSM techniques exploit the correlation between soil properties and soil forming factors by using remote sensing data as covariates. Geostatistical analysis has been one of the DSM techniques that has been widely applied (Biswas *et al.*, 2013;Veronesi *et al.*, 2012; Friedel and Iwashita, 2013).

Soil depth is a key variable for soil moisture storage capacity. It controls numerous surface, vadose processes and generally the hydrological response of a landscape (Lacoste *et al.*, 2016). Therefore, knowledge of soil MHC distribution in the vertical dimension and landscape is important. Lateral and vertical (3-dimensional) characterisation of spatial variability of soil MHCis warranted for prediction of un-saturated water and solute flow in the vadose (Saito *et al.*, 2009; Vereecken *et al.*, 2010).Numerous studies have coupled geostatistical analysis and soil depth functions to characterise the 3-dimensional variation of soil properties(Malone *et al.*, 2009; Veronesi *et al.*, 2012). Soil depth functions model discrete soil horizon data into a continuous distribution (Malone *et al.*, 2009; Minasny *et al.*, 2013). Equal area quadratic splines have been highlighted as the most efficient soil depth functions (Malone *et al.*, 2009; Adhikari *et al.*, 2013).

1.4 High resolution mapping of MHC in data sparseregions

High resolution MHC mapping for large spatial domains in data sparseregionsis a challenge (Odgers *et al.*, 2015; Malone *et al.*, 2016). There is need to develop approaches to produce more accurate, complete and consistent maps (Hengl *et al.*, 2017a). Though the idea ofsampling a sufficient MHC dataset appeals (Malone *et al.*, 2016), it is often impractical or prohibitive due to cost or timeconstraints

(Odgers *et al*., 2015). Therefore, there is need toconsider less costly approaches like model extrapolation (Malone *et al*., 2016).

The widely applied geostatistical analysis is highly sensitive to small data sets (Heuvelink, 2014). Further, geostatistical analysis assumes that data is normally distributed (Kavianpoor *et al*., 2012). This makes the geostatistical approach less ideal for mapping soil propertydata like MHC which is non-linear (Vereecken *et al*., 2010) andis of a non-normal distribution class. Machine learning algorithms (MLAs) are an attractivealternative for DSM ofsoil MHC (Kovačević*et al*., 2010; Ließ*et al*., 2016; Hengl *et al*., 2017a). MLAs are non-parametric (Ustuner *et al*., 2015) and make no assumption on data distribution class.

Machine learning algorithms are generally a broad set of models used to determine patterns in data and to make predictions (Brungard *et al*., 2015). They have gained popularityin DSM (Odgers *et al*., 2011; Brungard *et al*., 2015; Taghizadeh-Mehrjardi*et al*., 2015; Hengl *et al*., 2017a;Hengl *et al*., 2017b). Some MLAs applied in DSM are; artificial neural networks, support vector machines (Kovačević*et al*., 2010; Ließ*et al*., 2016), k-Nearest Neighbours (Taghizadeh-Mehrjardi*et al*., 2015). Random forests (RF) (Taghizadeh-Mehrjardi*et al*., 2015; Hengl *et al*., 2017a; Hengl *et al*., 2017b) and cubist decision trees (CB)(Adhikari *et al*., 2013; Adhikari *et al*., 2014) are more often applied in DSM.

Though MLAsare often applied to large datasetsfor soils mapping (Brungard *et al*., 2015; Hengl *et al*., 2017a; Hengl *et al*., 2017b), some case studies have highlighted positive results on small soil datasets (Khosi *et al*., 2016). Where sparse and spatially referenced discontinuous soil data exists, the soil data can be linked to spatial environmental covariates like remote sensing data, using MLAs to generate spatially continuous maps (Nussbaum *et al.*, 2017). With remote sensing data increasingly becoming freely available, fitting novel MLAs to large sets of covariates for soil mapping is now common (Ließ*et al.*, 2016; Hengl *et al.*, 2017a). However, relationships between soil properties like MHC and spatial environmental covariates need to be better understood (Brungard *et al.*, 2015). This research work was thus aimed at improving understanding on estimation of soil moisture holding capacity at large spatial domains using machine learning algorithms.

1.5 Justification

Crop production is currently facing unprecedented challenges inter alia climate change and increasing water scarcity (Raes *et al.*, 2009; Huang and Li, 2010). Establishing and sustaining resilient agricultural systems that can optimally utilise the scarce water resources for sustainable crop production will be a key mitigation strategy. The mechanism involves an assessment of the impact of agricultural practices and climate on resource use dynamics and crop yields. Simulation models (e.g. crop models such as AQUACROP, DSSAT and APSIM) with soil moisture holding properties as input parameters are key components of impact assessments to predict the hydrological, ecological or economic effects of the prospective climate change on agricultural production (Raes *et al.*, 2009). This research contributes to improvement of the predictive capacity and utility of simulation models, through better scaling of soil MHC at higher system levels. The research will provide a spatially-explicit decision aid to support policy formulations for crop production management at different scales with direct benefit in precision agriculture, irrigation water management, and development of agricultural production zones.

1.2 Objectives of the Study

The general objective of this research was to improve understandingon estimation of soil moisture holding capacity at large spatial domains using machine learning algorithms.

Specific objectives were to:

i. Assess the suitability of a probabilistic sampling scheme for mapping soil moisture holding capacity.

ii. Evaluate pedo-transfer functions for prediction of soil moisture holding capacity.

iii. Evaluate random forests and cubist decision trees for spatial estimation of soil moisture holding.

1.2.1 Hypotheses

The following hypotheses were tested:

1) Probabilistic and non-probabilistic sampling yield similar spatial estimates

2) Mean MHC predictions for SVM PTFs and MLR-PTFs are equal

3) Random forests and cubist decision trees generate similar spatial estimates of MHC

4) Combining measured and predicted MHC data in DSM yields similar accuracy in MHC maps like using measured data alone

1.3 Outline ofDissertation

The research in this dissertation was aimed at improving understanding on estimation of soil moisture holding capacity at large spatial domains using machine learning algorithms. Chapter One is a general introduction highlighting the research gaps. Chapters Two to Four are manuscripts discussing approaches contributing to fulfilling the main research objective. Chapter Two assesses suitability of a probabilistic sampling scheme for mapping soil moisture holding capacity.A k-means algorithm was used for establishing two sampling designs – stratified random sampling and spatial coverage sampling. Geostatistical analyses were performed to assess the two designs for mapping MHC. Chapter three addresses the development of PTFs for estimation of soil moisture holding capacity. Support vector machine learning algorithm was used for developing predictive models for FC and WP. Comparisons were made with multiple linear regression method. Chapter Fourevaluates two machine learning algorithms for3-Dimensional mapping of soil moisture holding capacity. An evaluation of the Random Forests and Cubist machine learning algorithms for predictive mapping was performed in combination with soil depth functions. Mapping accuracy was also evaluated with substitution of measured data with PTF-estimated data. Chapter Five is a synthesis of the research work with general conclusions and recommendations.

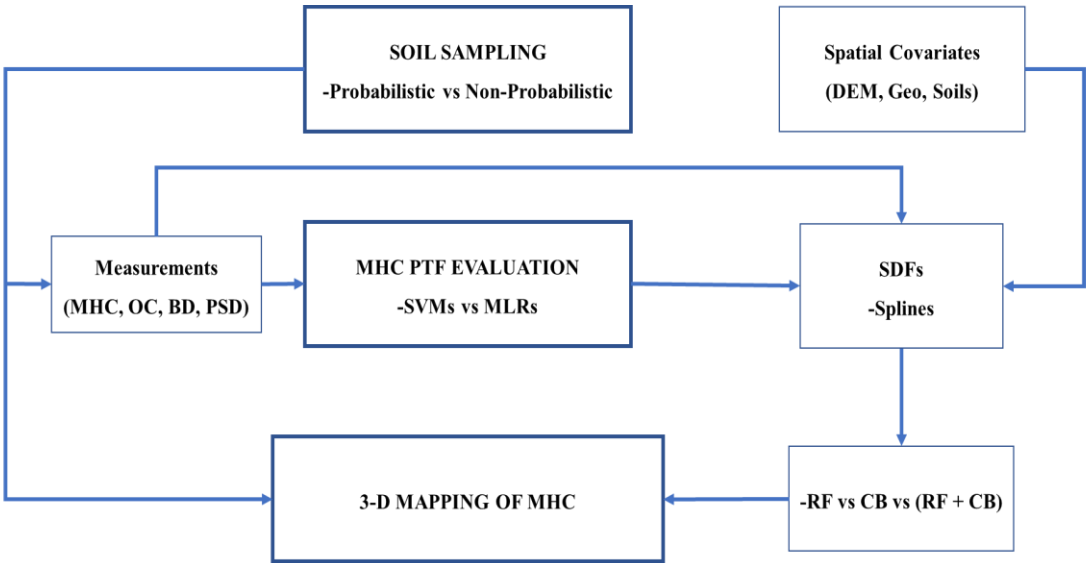

Figure 1.1: Conceptual Framework

1.4 REFERENCES

Adhikari, K., Bou Kheir, R., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A. B. and Greve, M. H. (2013). High-Resolution 3-D Mapping of Soil Texture in Denmark. *Soil Science Society of America Journal* 77(3): 860 – 876.

Adhikari, K., Hartemink, A. E., Minasny, B., Bou Kheir, R., Greve, M. B., and Greve, M. H. (2014). Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark. *PLoS ONE* 9(8): e105519.

Asgarzadeh, H., Mosaddeghi M. R., Mahboubi, A. A., Nosrati, A. and Dexter, A. R. (2010). Soil water availability for plants as quantified by conventional available water, least limiting water range and integral water capacity. *Plant Soil* 335:229 – 244.

Barron, J. (2004). Dry spell mitigation to upgrade semi-arid rainfed agriculture: Water harvesting and soil nutrient management for smallholder maize cultivation in Machakos, Kenya. Dissertation submitted to Stockholm University for award of PhD Degree, Stockholm, Sweden. 39pp.

Biswas, A., Cresswell, H. P.,Chau, H. W., Rossel, R. A. V., Si, B. C. (2013). Separating scale-specific soil spatial variability: A comparison ofmulti-resolution analysis and empirical mode decomposition. *Geoderma* 209–210: 57–64

Bou Kheir, R., Bøcher, P. K., Greve, M. B. and Greve, M. H. (2010). The application of GIS based decision-tree models for generating the spatial distribution of hydromorphic organic landscapes in relation to digital terrain data. *Hydrology and Earth System Sciences*14: 847 – 857.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., Edwards Jr., T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma* 239–240: 68–83

Brus, D. J. and Heuvelink, G. B. M. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* 138: 86–95.

Brus, D. J., Kempen, B., Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science* 62: 394 – 407.

Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J., Smaling, E.M.A. (2014). Soil organic carbon stocks in the Limpopo National Park, Mozambique: Amount, spatial distribution and uncertainty. *Geoderma* 213 : 46–56

Chirico, G. B., Medina, H. and Romano, N. (2007). Uncertainty in predicting soil hydraulic properties at the hillslope scale with indirect methods. *Journal of Hydrology* 334: 405 – 422.

Enfors, E. I., Gordon, L. J., Peterson, G. D. and Bossio, D. (2008). Making Investments in Dryland Development Work: Participatory Scenario Planning in the Makanya Catchment, Tanzania.*Ecology and Society* 13(2): 42

Friedel, M. J. and Iwashita, F. (2013). Hybrid modeling of spatial continuity for application to numerical inverse problems. *Environmental Modelling and Software* 43: 60 – 79.

Gaiser, T., Graef, F. and Cordeiro, J. C. (2000). Water retention characteristics ofsoils with contrasting clay mineral composition in semi-arid tropical regions.Australian Journal of Soil Resources 38: 523 - 536.

Garrity, D., Dixon, J., and Boffa J-M. (2012). Understanding African Farming SystemsScience and Policy Implications. Paper presented atFood Security in Africa: Bridging Research and Practice Conference, Sydney, Australia. 29th – 30th November 2012. 55pp.

Haghverdi, A., Cornelis, W. M. and Ghahraman, B. (2012). A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. *Journal of Hydrology* 442 – 443: 46 – 54.

Hengl T, Heuvelink GB, Kempen B, Leenaars JG, Walsh MG,Shepherd KD, Sila A, MacMillan RA, Mendes de Jesus J,Tamene L, Tondoh JE (2015) Mapping soil properties ofAfrica at 250 m resolution: random forests significantlyimprove current predictions. *PLoS ONE* 10:e0125814.

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. (2017a). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* 12(2): e0169748

Hengl, T., Leenaars, J. G.,Shepherd K. D., Walsh M. G.,Heuvelink, G. B. M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E.,Wheeler, I.,and Kwabena, N. A. (2017b). Soil nutrient maps of Sub-Saharan Africa: assessment of soilnutrient content at 250 m spatial resolution using machinelearning. *Nutrient Cycling in Agroecosystems* 109:77–102.

Heuvelink, G. B. M. (2014). Uncertainty quantification of GlobalSoilMap products. In: *GlobalSoilMap: basis of the global spatial soil information system. Proceedings of 1st GlobalSoilMap Conference*(Edited by Arrouays, D.et al.).7-9 October 2013. Orleans, France. CRC Press. Leiden – Netherlands. pp. 335-340

Huang, F. and Li, B. (2010) Assessing grain crop water productivity of China using a hydro-model-coupled-statistics approach Part I: Method development and validation. *Agricultural Water Management* 97: 1077 – 1092.

Kavianpoor, A., Ouri, A. E., Jafarian Jeloudar, Z., Kavian, A. (2012). Spatial Variability of Some Chemical and Physical Soil Properties in Nesho Mountainous Rangelands. *American Journal of Environmental Engineering* 2(1): 34-44

Kempen, (2011). Updating soil information with digital soil mapping. Thesis submitted for award of PhD Degree of Wageningen University, Wagenigen - Netherlands. 218pp.

Khlosi, M., Alhamdoosh, M., Douaik, A. Gabriels, D. and Cornelis, W. M. (2016). Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. *European Journal of Soil Science* 67: 276 - 284

Kilasara, M. (2010). Selection and use of soil characteristics in digital soil mapping in Tanzania. In: *Proceedings of the 19th World Congress of Soil Science*.(Edited byGilkes, R. and Prakongkep,N.)1 – 6 August 2010, Brisbane, Australia. 377 - 378pp.

Kovačević, M., Bajat, B., Gajić, B. (2010). Soil type classification and estimation of soil properties using support vector machines. *Geoderma* 154: 340 - 347

Lacoste, M., Mulder, V.L., Richer-de-Forges, A. C., Martin, M.P., Arrouays D. (2016). Evaluating large-extent spatial modeling approaches: A case study forsoil depth for France. *Geoderma Regional* 7: 137–152.

Ladson, A. R., Lander, J. R., Western, A. W., Grayson, R. B. and Zhang, L. (2006).Estimating extractable soil moisture content for Australian soils from fieldmeasurements. *Australian Journal of Soil Research* 44: 531 - 541.

Lagacherie, P., Anne-Ruth Sneep, A. R., Cécile Gomez, C., Sinan Bacha, S., Coulouma, G., Mohamed Hédi Hamrouni, M. H., Mekki, I. (2013). Combining Vis–NIR hyperspectral imagery and legacy measured soilprofiles to map subsurface soil properties in a Mediterranean area(Cap-Bon, Tunisia).*Geoderma* 209–210 :168–176.

Lark, R. M. (2016). Changes in the variance of a soil property along a transect, a comparisonof a non-stationary linear mixed model and a wavelet transform.*Geoderma* 266: 84–97

Leenaars, J. G.B., Hengl, T., Gonzalez, M. R., de Jesus, J. M. Heuvelink, GBM, Wolf, J., van Bussel, L., Claessens, L., Yang, H. and Cassman, K. G. (2015). *Root zone plant-available water holding capacity of the Sub-Saharan Africa soil, version 1.0. Gridded functional soil information. (dataset RZ-PAWHC SSA v 1.0).* ISRIC Report 2015/02: Collaboration project of Africa Soil Information System and Global Yield Gap and Water Productivity Atlas (GYGA). ISRIC- World Soil Information, Wageningen, the Netherlands. 108pp.

Levi, M. R., Schaap, M. G. and Rasmussen, C. (2015). Application of Spatial Pedotransfer Functions to Understand Soil Modulation of Vegetation Response to Climate. *Vadose Zone Journal* 14 (9): 1-14.

Ließ, M., Schmidt, J., Glaser,B. (2016). Improving the Spatial Prediction of Soil Organic Carbon Stocks in a Complex Tropical Mountain Landscape by Methodological Specifications in Machine Learning Approaches. *PLoS ONE* 11(4): e0153673.

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

Malone, B. P., Jha, S. K., Minasny, B., McBratney, A. B. (2016). Comparing regression-based digital soil mapping and multiple-pointgeostatistics for the spatial extrapolation of soil data. *Geoderma* 262: 243–253.

Malone, B. P., Styc, Q.,Minasny, B., McBratney, A. B. (2017). Digital soil mapping of soil carbon at the farm scale: A spatialdownscaling approach in consideration of measured and uncertain data. *Geoderma* 290: 91–99.

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

Minasny, B., Whelan, B. M., Triantafilis, J. and McBratney A. B. (2013). Pedometrics research in the vadose zone — Review and Perspectives. *Vadose Zone Journal*12(4): 1-20.

Moore, I. D., Gessler, P. E., Nielsen, G. A. and Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* 57: 443 – 452.

Nussbaum, M., Walthert, L., Fraefel, M., Greiner, L. and Papritz, A. (2017). Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models. *Soil Discussions.* (In press).

Odgers, N., McBratney, A. B. Minasny, B. (2011). Bottom-up digital soil mapping. I. Soil layer classes. *Geoderma* 163: 38–44

Odgers, N., McBratney, A. B. Minasny, B. (2015). Digital soil property mapping and uncertainty estimation using soil class probability rasters. *Geoderma* 237–238: 190–198

Pender, J., Place, F., Ehui, S. (1999). Strategy for agricultural development in the East African highlands. Environment and Production Technology

Division – International Food Policy Research Institute. EPTD Discussion Paper No. 41. 86pp.

Raes, D., Steduto, P., Hsiao, T. C. and Fereres, E. (2009). AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: II. Main Algorithms and Software Description. *Agronomy Journal* 101(3): 438 – 447.

Reynolds, T. W., Waddington, S. R., Anderson, C. L., Chew, A., True, Z. and Cullen,A. (2015). Environmental impacts and constraints associated with the production of major food crops in sub-Saharan Africa and South Asia. *Food Sec.* 7:795–822

Rizzo, R., Demattê, J. A. M., Igo F. Lepsch, I. F., Gallo, B. C., Caio T. Fongaro, C. T. (2016). Digital soil mapping at local scale using a multi-depth Vis–NIR spectral library and terrain attributes. *Geoderma* 274: 18–27

Rockstrom, J., Hatibu, N., Oweis, T. Y., Wani, S., Barron, J., ¨ Bruggeman, A., Farahani, J., Karlberg, L., and Qiang, Z. (2007). Managing water in rainfed agriculture. In: *Water for food, water for life: A comprehensive assessment of water management in agriculture*. (Edited by: Molden, D.*et al.*) Earthscan, London. pp. 315–352

Saito, H., Seki, K. and Simunek,J. (2009). An alternative deterministic method for the spatial interpolation of water retention parameters. *Hydrology Earth System Sciences* 13: 453 - 465.

Santra, P., Chopra, U. K. and Chakraborty, D. (2008). Spatial variability of soil properties and its application in predicting surface map of hydraulic parameters in an agricultural farm. *Current Science* 95(7): 937 - 945.

Santra, P., Sahoo, R. N., Das, B. S., Samal, R. N., Pattanaik, A. K., Gupta, V. K. (2009). Estimation of soil hydraulic properties using proximal spectral reflectance in visible, near-infrared, and shortwave-infrared (VIS–NIR–SWIR) region. *Geoderma* 152: 338 – 349.

Shiferaw, B., Tesfaye, K., Kassie, M., Abate,T., Prasanna, B.M., Menkir, A.(2014). Managing vulnerabilitytodroughtandenhancinglivelihoodresiliencein sub-SaharanAfrica:Technological,institutionalandpolicyoptions. *Weather and Climate Extremes* 3: 67–79.

Si, B. C. (2008). Spatial Scaling Analyses of Soil Physical Properties: A Review of Spectral and Wavelet Methods. *Vadose Zone Journal* 7: 547 – 562.

Taghizadeh-Mehrjardi, R. Nabiollahi, K., Minasny, B., Triantafilis,J. (2015). Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran.*Geoderma* 253–254: 67–77

Ustuner, M., Sanli F. B. and Dixon, B. (2015) Application of Support Vector Machines for Landuse Classification Using High-Resolution RapidEye Images: A Sensitivity Analysis. *European Journal of Remote Sensing* 48: 403.-.422.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M. G. and van Genuchten, M.Th. (2010). Using Pedotransfer Functions to Estimate the van Genuchten– Mualem Soil Hydraulic Properties: A Review. *Vadose Zone Journal* 9: 795 – 820.

Veronesi, F., Corstanje, R. and Mayr, T. (2012). Mapping soil compaction in 3D with depth functions. *Soil & Tillage Research* 124:111 – 118.

Xu, X., Kiely, G. and Lewis, C. (2009). Estimation and analysis of soil hydraulic properties through infiltration experiments: comparison of BEST and DL fitting methods. *Soil Use and Management* 25: 354 – 361.

Walvoort, D.J.J., Brus, D.J., de Gruijter, J. J. (2010). An R-package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences* 36: 1261–1267

Wang, L., D'Odorico,P., Evans, J. P., Eldridge, D. J., McCabe, M. F.Caylor, K. K., and King,E. G. (2012). Dryland ecohydrology and climate change: Critical issues andtechnical advances. Hydrol. Earth System Science 16: 2585–2603

Zhu, J. and Mohanty, B. P. (2006). Effective scaling factor for transient infiltration in heterogeneous soils. *Journal of Hydrology* 319: 96 – 108.

Zinn, Y. L., Lal, R. and Resck, D. V. S. (2005). Texture and organic carbon relations described by a profile pedotransfer function for Brazilian Cerrado soils. *Geoderma* 127: 168– 173.

CHAPTER TWO

2.0 OPTIMISING SAMPLING DESIGN FOR MAPPING SOIL MOISTURE HOLDING CAPACITY USING PROBABILISTIC APPROACH

ABSTRACT

Management of moisture in cropping systems necessitates a characterisation of spatial variability of soil moisture holding capacity (MHC). Geostatistical methods as part of the digital soil mapping (DSM) toolset provide a statistical approach for scaling the variability of soil MHC. Probabilistic and uniform geographic distribution of the sampling points over the target area provides unbiased and high gains in accuracy of predictions of MHC. The main objective of this study was, therefore, to optimise a probabilistic sampling scheme for mapping MHC using a *k*-means clustering algorithm. This was implemented in R-software using the *spcosa* package.The study area was divided into 100 geostrata of equal area and sample locations randomly selected from each geostratum to establish a stratified random sampling scheme (STRS). Soil samples were taken at each of the sampling locations and analysed.The STRS method was compared to a spatial coverage sampling scheme (SPCS) using variography analysis and ordinary kriging interpolation with the R-software *gstat* package. Spatial dependence of STRS was 72 %, indicating a weaker spatial structure than forSPCS with a spatial dependence of 19 %. Performance indicators for ordinary kriging (Mean Error (ME) and Root Mean Square Error (RMSE)), respectively, were slightly better for STRS (0.0014 and 0.156) than SPCS (0.0017 and 0.161). The STRS is therefore an optimal sampling scheme for mapping the spatial distribution of soil MHC.

Keywords: Moisture holding capacity, digital soil mapping, sampling, *k*-means, geostrata

2.1 INTRODUCTION

Availability of soil moisture in crop production systems in sub-Saharan Africa is a widespread challenge for sustainable crop production in smallholder farms (Barron, 2004). Knowledge of soil moisture storage conditions is important for management and continuity of moisture supply within these agricultural areas (Santra *et al.*, 2008). Moisture storage capacity of soils is quantified by moisture holding capacity (MHC) which is defined as the difference between moisturecontent at field capacity (FC) and wilting point (WP) (Schoonover and Crim, 2015; Asgarzadeh*et al.*, 2010). Field capacity is characterised as moisture retained in a saturated soil under free drainage for a period of 48 hours(Novák and Havrila, 2006).A soil matric suction of 33 kPahasbeen widely applied as theupper limit of MHC(i.e. FC) (Gaiser *et al.*, 2000). Wilting point is defined as soil moisture content after equilibration at a matric suction of 1500 kPa. Wilting point is considered as soil moisture content at which plant'sturgidity cannot recover even with additional replenishment of moisture to the soil matrix (Asgarzadeh*et al.*, 2010).

Moisture holding capacity has found wide applications as an input in crop growth simulation models (Vereecken *et al.*, 2008; Gaiser *et al.*, 2000)for yield potential or yield gap prediction, irrigation planning (Santra *et al.*, 2008), land suitability modelling, and food security monitoring applications. Input MHC data for running these processes is seldom in sufficient supply or in appropriate format for the aforementioned tasks (Gaiser *et al.*, 2000). Conventionally, data is collected from point locations and represented as discrete choropleth maps for areas attributable to similar soil characteristics. However, this approach involves large uncertainties as soil variables intrinsically exhibit high spatial heterogeneity (Heuvelink and Webster, 2001; Santra *et al.*, 2008; Odgers *et al.*, 2011).

An account of associated data uncertainties is fitting for decision support on judicious exploitation and management of soil resources (Heuvelink *et al.* 2007).Probabilistic methods offer a subtle cue for mapping the stochasticity associated with the spatial distribution of soil MHC. Geostatistics which is part of the toolsets in digital soil mapping (DSM) provides means for mapping the spatial distribution of MHC (Minasny and Hartemink, 2011; Cambule *et al.*, 2013). Map quality is a function of the configuration of the sampling pattern (Brus and Heuvelink, 2007; Marques Jr *et al.*, 2015). High gains in prediction accuracy necessitate regularly placed and uniformly distributed sampling locations across the geographic space (Walvoort *et al.*, 2010). A systematic regular sampling grid has been a widely preferential scheme for uniform distribution of sampling points over the target geographic space. It is, however, restrictive where patterns of periodicity associated with the underlying topography, land use or geology are inherent within the area. Unbiased data on the statistical distribution of soil variables is useful for reporting associated uncertainty estimates (Brus *et al.*, 2011; Webster and Lark, 2013). This calls for randomisation of sampling locations within the target geographic space. Randomisation of sampling locations enables computation of the probability distributions from which confidence limits can be assessed and also provides a means of setting local reference values such as what is low, moderate or high.

Several sampling algorithms have been proposed over the years for optimising random sampling nodes over the geographic space. Among these are spatial simulated annealing (Szatmári *et al*., 2015), conditioned Latin hypercube sampling (Minasny and McBratney, 2006), k-means sampling (Walvoort *et al.,* 2010), and most recently the balanced sampling design (Brus, 2015). These methods differ in The objective function for optimisation of sampling locations differs among these methods. A number of them, however, depend on the minimisation of the kriging variance as the objective criterion. This calls for the estimation of a variogram of the variable of interest which is not feasible in previously unsampled localities (Brus and Heuvelink, 2007). The k-means sampling algorithm proposed and implemented in Walvoort *et al.* (2010), uses the mean squared shortest distance between fine grid cells of a discretised target area as the objective function for the optimisation of sampling locations. The co-ordinates of the midpoints of these grid cells are the classification variables. A key advantage of the k-means approach is the elimination of the need for priori variogram estimation and provides an approach for stratification of the geographic space. The objective of this study was therefore to develop a probabilistic sampling scheme for mapping soil moisture holding capacity using k-means clustering and assess its suitability.

2.2 MATERIALS AND METHODS

2.2.1 Study area

The study area was Ilakala village in Kilosa District - Tanzania. It is located within latitudes 7^{o} 6’ 30” S and 7^{o} 9’ 30” S, longitudes 36^{o} 51’ 30” E and 36^{o} 57’ 30” E; with an area of about 44 km^{2} (Fig. 2.1). It borders Mikumi National Park to the south. Altitude ranges from 514 m to 896 m with a hilly relief forming part of streak of the Eastern Arc mountains straddling the South Western and Western fringes, predominantly covered byMiombo woodlands. Southern areas are predominantly covered byarable agriculture with some pastoral communities settled within the area. The area has a dry sub-humid climate with a unimodal crop growing season andmean annual rainfall of about 500-800 mm. The major crops grown within the area are sesame, maize and pigeon peas.

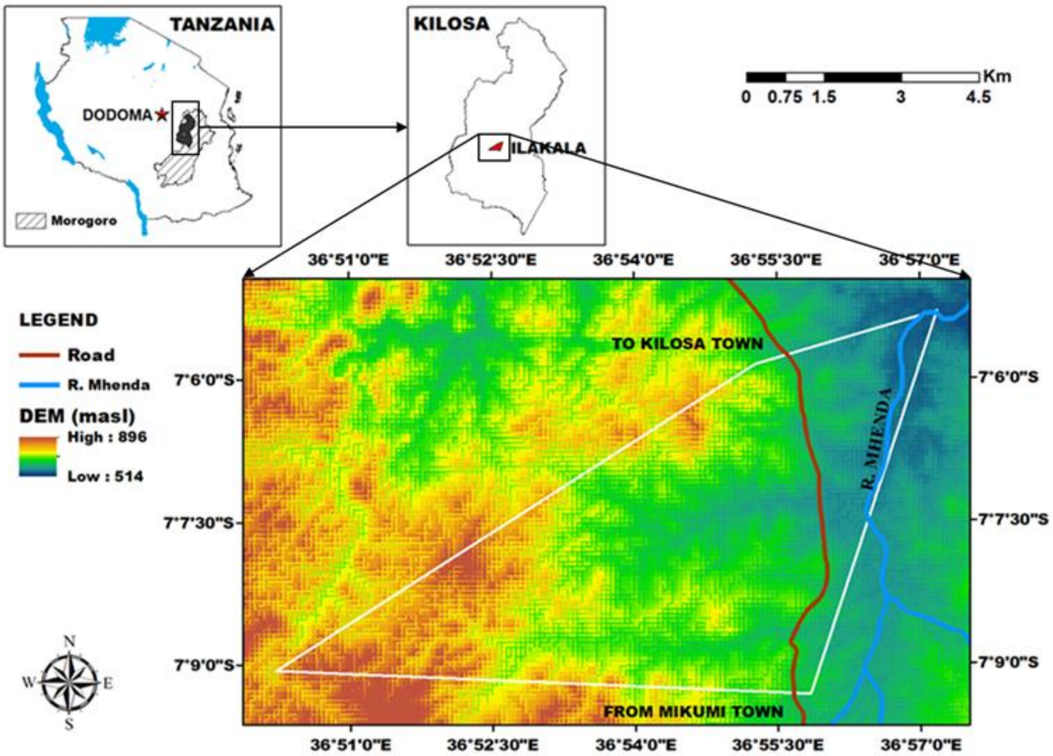

Figure 2.1: Study area with the soil sampling locations

2.2.2 Establishingsampling scheme

The study was designed to establish a sampling scheme of 100 evenly distributed spatial points. The decision to select 100 sampling points was informed by the financial outlay available to execute the study.Robinson and Metternicht (2006) highlighted that a minimum of 100 – 150 data points was necessary to achieve a stable semivariogram. Brungard and Boettinger (2010) found thata sample size of about 200-300 was optimal for digital soil mapping in an area of 300 km^2. Therefore, the quota of sampling points in this study is satisfactory for assessment of spatial structure for DSM.

The R-package *spcosa* (Walvoort *et al.*, 2010) was used for optimising the distribution of 100 sampling points. A shapefile of the study area was imported into the R-software environment and discretised into 100 equal geographic strata (geostrata). Three sampling locations in each stratum (geostratum) were established using a stratified random sampling approach. One of the points was the primary sampling point and the other additional two points as ordered contingency sampling locations. The order of selection of the three sampling points within a geostratum was strictly followed.Sampling points would be visited during the field campaign in order i.e. the primary sampling point, then first contingency point and ultimately the second contingency point. Contingency points would only be visited for sampling when the primary sampling point appeared unsuitable for sampling. Reasons for not sampling a selected random point were for instance, denial of access or non-availability of soil (falling in a rocky area or water body). It was not permissible within the design to select the sampling point closest to the primary point once this initial point appeared unsuitable for sampling.For comparison purposes, a spatial coverage sampling design was also established using the centroids of the geostrata as the sampling nodes.

2.2.3 Soil sampling and analysis

Bulk soil samples of about 0.5 kg were taken from 100 sampling nodes in a stratified random sampling scheme at three depth intervals of 0 -30, 30 – 60 and 60 - 100 cm. The bulk soil samples were air dried and crushed and sieved through a 2 mm sieve. Sieved soil samples were then analysed in the laboratory for particle size distributionand organic carbon. Particle size fractions were determined by the Bouyoucos hydrometer method (Gee and Bauder, 1986) and separated according to the United States Department of Agriculture particle size classification system (FAO, 2006). Organic carbon was determined by the wet oxidation method of Walkley and Black (Nelson and Sommers, 1982). Undisturbed soil core samples in 100 cc Kopecky rings with height and diameter dimensions of 5 cm were used to determine soil moisture at -30 kPa and -1500 kPa with a pressure plate apparatus. The very soil core samples were used to determined bulk density after drying the soil core samples at 105^{o} C for 24 hours(Blake and Hartge, 1986). A range of matric suctions for FC have been reported in literature ranging from 5 kPa to 33 kPa (Asgarzadeh*et al.*, 2010). For this study FC was considered as soil moisture at -30 KPa matric suction. Therefore, MHC was calculated as the difference between moisture content at FC (-30 KPa) and WP (-1500 KPa).

2.2.4 Statistical data analyses

All exploratory statistics of the measured soil properties (minimum, maximum, mean, standard deviation, kurtosis and skewness) were analysed in R-statistical software(R Core Team, 2016).

2.2.5 Analysis of spatial structure and interpolation of MHC

Spatial structure of soil MHC data wasanalysed through (semi-)variogram modelling. Variogram modelling represents spatial variability of MHC data atsampling locations as a function of theirseparation distance (referred to as the lag) in a graph known as a semivariogram (Robinson and Metternicht, 2006; Diggle and Ribeiro Jr., 2007). Ordinary kriging (OK) technique was used for spatial interpolation of MHC data (Diggle and Ribeiro Jr., 2007). OK is an optimum interpolation technique that quantifies unbiased linear estimatesof regionalized variables at unsampled locationswith the aid of the structural properties of the semivariogramand the initial set of data at sampled locations(Chilesand Delfiner, 1999; Huang *et al.*, 2006; Dumitrescu *et al.*, 2015).Analysis of spatial structure and OK interpolation of the distribution of MHC was performed usingR-software *gstat* package (Pebesma, 2004).

Theempirical semivariogram for MHC was computed using Equation 1(Chilesand Delfiner, 1999). Computation of the empirical semivariogram results in an uneven scatter unsuitable for calculating the OK weights.Therefore, a mathematical model referred to as a ‘theoretical model of semivariogram’ was fitted to the empirical semivariogram to derive three essential parameters for deriving the OK weights – the nugget, sill and range. The nugget defines small-scale variation and measurement error within MHC data. Partial sill indicates the amount of variation represented by the spatial correlation structure (Santra *et al.*, 2008). The partial sill increases with increasing lag until a lag value where the partial sill equals the variance of the data. This lag value is known as the range. The range is the maximum separation distance (lag) at which spatial autocorrelation exists between any two sampling points.

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [\, Z(s_i) - Z(s_i + h)\,]^2 \tag{1}$$

Where *Z(*s_i*)* is the measured MHC value at sampled location $s_i$with coordinatevector (x_i, y_i), *Z(*s_i*+h)*is the MHC value at a lag of h from location s_i,*N*(*h*) are the sample pairs within a lag interval of h.

Residual maximum likelihood (REML) procedure was applied to fit a theoretical variogram model to the empirical variogram (Pebesma, 2004). Arbitrary estimates of the theoretical variogram parameters (nugget, partial sill and range) were selected to initialise this REML procedure. The resultant nugget, sill and range parameters from the fit were used in a subsequent step to calculate semivariances of the MHC data. Semivariances were utilised in the ordinary kriging system for generation of the MHC prediction map. Leave one out cross validation (LOOCV) was utilised to assess the performance of sampling schemes. In LOOCV approach, data for a target prediction location was eliminated and the remaining dataset was used for prediction of the MHC value at that location (Robinson and Metternicht, 2006; Li *et al.*, 2012). This processwas repeated for each sampling location andperformance indices were then computed from the pool of paired actual measured MHC value and the resultant LOOCV predicted MHC value at each location. Performance indices used in the evaluation of ordinary kriging predictions were the mean error (ME) (Eq. 2), root mean square error (RMSE) (Eq. 3), average kriging standard error (ASE) (Eq. 4), and root mean square standard error (RMSSE) (Eq. 5).

$$\mathbf{ME} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}[\,\boldsymbol{Z}(\mathbf{x_i}) - \mathbf{Z}^*(\mathbf{x_i})\,] \tag{2}$$

$$\mathbf{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}[\,\boldsymbol{Z}(\boldsymbol{x_i}) - \boldsymbol{Z}^*(\boldsymbol{x_i})\,]^{\mathbf{2}}} \tag{3}$$

$$\mathbf{ASE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{\sigma}^{\mathbf{2}}(\boldsymbol{x_i})} \tag{4}$$

$$\mathbf{RMSSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left[\frac{\boldsymbol{Z}(\boldsymbol{x_i}) - \boldsymbol{Z}^*(\boldsymbol{x_i})}{\boldsymbol{\sigma}^{\mathbf{2}}(\boldsymbol{x_i})}\right]^{\mathbf{2}}} \tag{5}$$

Where $Z(x_i)$ is the measured MHC value at location x_i,$Z^*(x_i)$ is the predicted MHC value at location x_i,n is thenumber of sampling points, and $\sigma^2(x_i)$ is the kriging variance for locationx_i.

2.3.1 Evaluation of sampling scheme

Figure 2.2A shows the geostrata developed by the k-means clustering method. The geostrata were one hundred and each had an area coverage of 0.44 km^2. Figures 2.2B to 2.2D illustrate the stratified random sampling scheme (STRS), spatial coverage scheme (SPCS) and actual sampled points, respectively. The STRS is shown with the primary sampling point and the two contingency points in each geostratum (Fig. 2.2B).

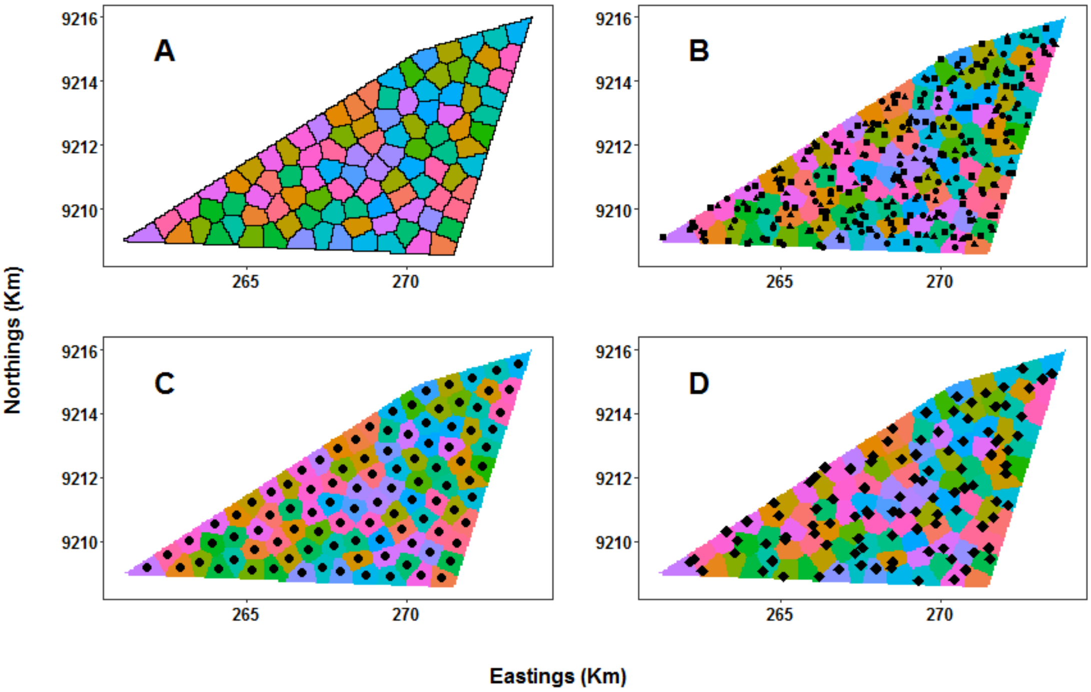

Figure 2.2: Geostrata (A), STRS (B), SPCS (C) and actual sampled points (D)

2.3.2 Descriptive statistics of soil properties

Measured MHC datasetfrom the 0-30 cm depth was used for analyses to compare the two sampling schemes with anassumption of stationarity across each geostratum (Diggle and Ribeiro Jr., 2007). Summary statistics of the soil properties for 0-30 cm depth are shown in Table 2.1. Moisture holding capacity (MHC) ranged from 0.02 to 0.1 cm^3 cm^{-3}. Mean moisture content values at field capacity (θ_{30}) and wilting point (θ_{1500}) were 0.37 cm^{3}cm^{-3} and 0.33 cm^{3}cm^{-3}, respectively. Bulk density (BD), clay, silt, and sand content ranged from 1.02 to 1.19 g cm^{-3}, 0.8 to 56.8 %, 2.8 to 27.8 %, and from 30 to 94.1 %, respectively. Organic carbon (OC) ranged from 0.16 to 3.37 %. Sand had the highest standard error of the mean while MHC had the lowest standard error of the mean. Figure 2.3 shows the textural class distribution of the soil dataset. The dataset consisted of a pool of seven different USDAtextural classes (Beaudette *et al.*, 2012), with most of the soil samples (> 55 %) of a sandy loam or a sandy clay loam textural class.

Table 2.1: Descriptive statistics of soil physicochemical properties

Variable	Minimum	Maximum	Mean (SE)	SD	Skewness	Kurtosis
MHC (cm^{3}cm^{-3})	0.02	0.10	0.04 (0.00)	0.02	1.43	2.13
θ_{30}	0.08	0.37	0.22 (0.01)	0.07	0.26	-0.66
θ_{1500}	0.06	0.33	0.18 (0.01)	0.07	0.34	-0.60
BD (gcm^{-3})	1.02	1.19	1.07 (0.00)	0.04	1.03	0.62
Clay (%)	0.80	56.80	18.94 (1.45)	14.51	0.78	-0.22
Silt (%)	2.80	27.80	14.16 (0.49)	4.92	-0.04	0.12
Sand (%)	30.00	94.10	66.9 (1.52)	15.18	-0.61	-0.29
OC (%)	0.16	3.37	1.02 (0.07)	0.70	1.29	1.54

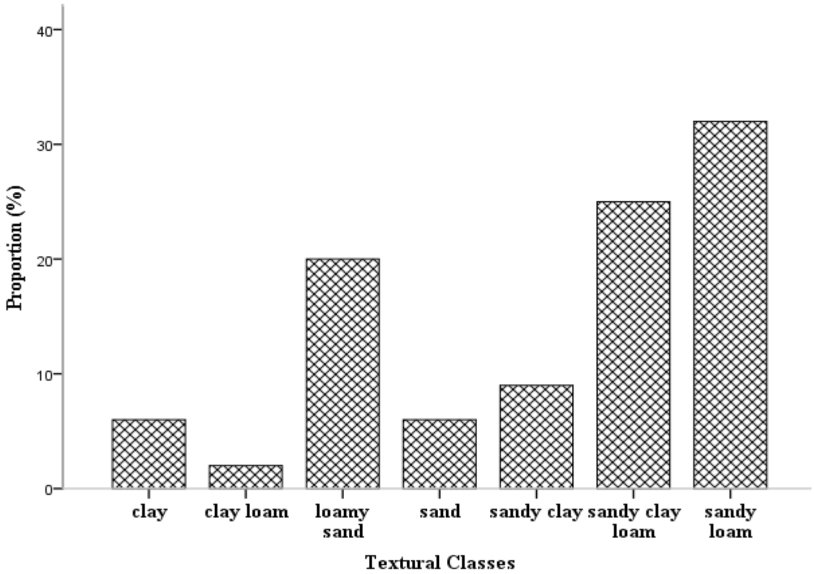

Figure 2.3: Textural class distribution of dataset

Skewness and kurtosis scores indicate that all the measured soil variables did not conform to a symmetrical Gaussian distribution (Table 2.1). Threshold values of skewness for a Gaussian distribution range from-0.391to 0.391 for a sample size of 100 (Doane and Seward, 2011). Though the skewness values for θ_{30}(0.26) and θ_{1500}(0.34) suggested a fit to the Gaussian distribution, the kurtosis values were negative indicating aplatykurtic distribution and thus deviating from the Gaussian distribution (De Carlo, 1997). Variable that have a normal distribution should ideally have a kurtosis of zero (0) (De Carlo, 1997).A histogram and probability density plot for both θ_{30} and θ_{1500} revealed that the two parameters had a bimodal distribution (Fig. 2.4A and Fig. 2.4B): a typical distribution with negative kurtosis (De Carlo, 1997). Moisture holding capacity data was leptokurtic with a strong positive skew anda long tail (Fig. 2.4C). The MHC data was log transformed (base 10) to generate data with a mild conformation to the Gaussian assumptions (Fig. 2.4D). This log transformed MHC was used for analysis of spatial structure and geostatistical interpolation.

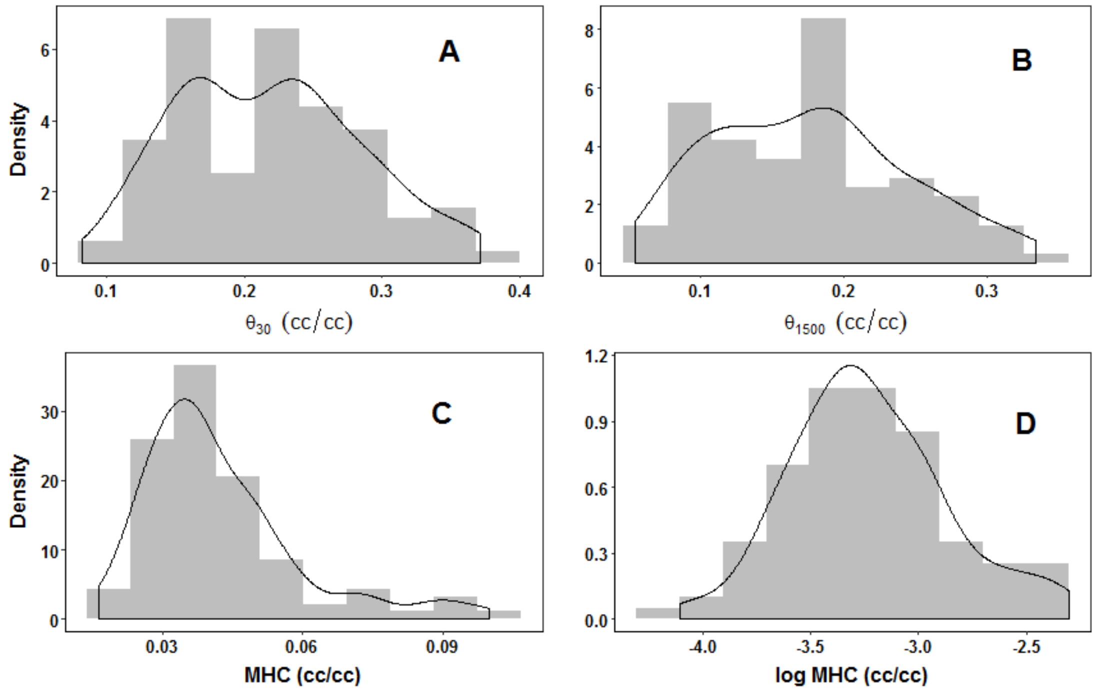

Figure 2.4: Histogram of moisture contents for FC (A), WP (B), MHC (C) and log transformed MHC (D)

2.3.3 Analysis of spatial structure and interpolation of MHC

Figure 2.5 shows the directional plots for MHC for the stratified random sampling scheme (STRS) and the spatial coverage sampling scheme (SPCS). Anisotropy (directional influence) was evident in the West to East direction (Fig. 2.5A and 2.5B) and to a lesser extent in South to North direction (Fig. 2.5C and 2.5D). The West-East trend seemed to fit a quartic polynomial while South-North trend seemed to fit a 3rd order polynomial.

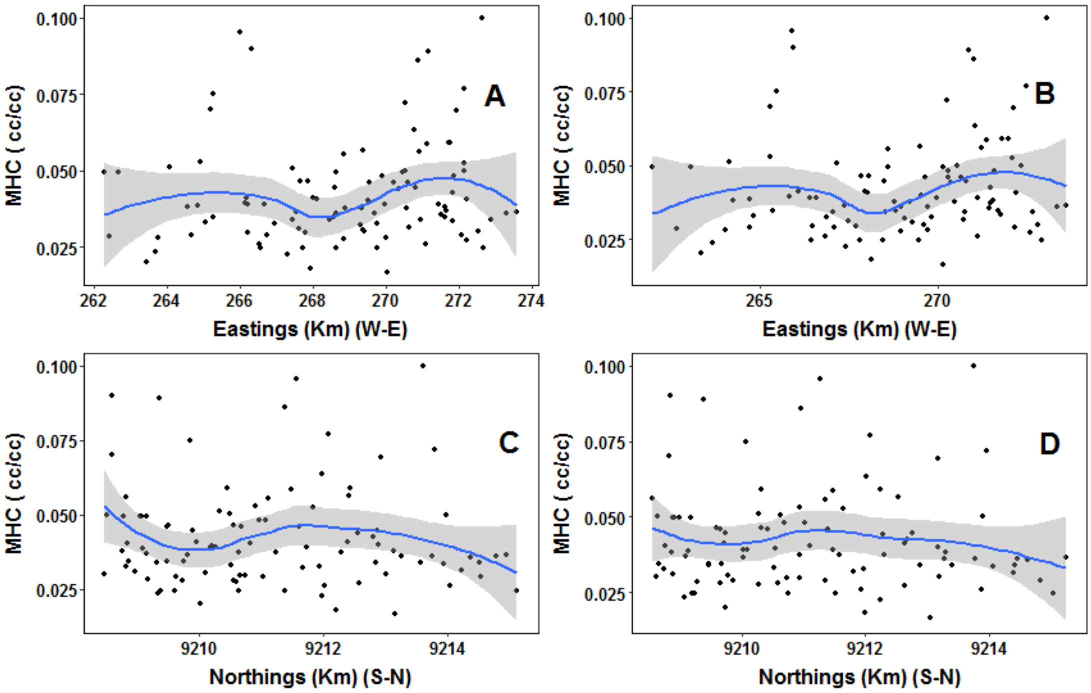

Figure 2.5: Directional plots of MHC for STRS (A & C) and SPCS (B&D)

Reasons for this trend were not straight forward but landscape characteristics (soil characteristics, land use and topography) seemed to offer the most plausible explanation for the anisotropy. The broad distribution of the land use pattern within the study area was such that forested landscapes predominate the South West and Western fringes of the study area while arable agriculture was prominent in a gradient to the orthogonal direction. Similarly, a topographic effect was evident with dissected hilly terrain along the South West to Western fringes and undulating topography in the opposite direction. Anthropogenic land use and topography are known to influence the development, variations in characteristics and spatial distribution of soil properties (Cambule *et al.*, 2013; Peng *et al.*, 2013). Land uses which lead to an accumulation of high organic matter content often result in higher moisture contents. Forest areas and some agricultural practices often fall in this category of land use.

Another posit is that the anisotropic effect could have been due to the geometric orientation of the study area. Diggle and Ribeiro Jr. (2007) aver that differential stretching and rotation of coordinate axes transform stationarity structure leading to geometrical anisotropy. The quadrangular planar geometry of the study area was longer in the South West to North East direction than the North-West to South East direction (Fig. 2.2A). The number of sampling locations increased gradually on the plane from the South East to North-West gradient across its orthogonal axis (Fig.2.2C andFig.2.2D). Periodicity associated with area size is apparent and influences the proportion of sampling locations on South West to North East axis. It is conceivable that fewer sampling locations are feasible towards the narrow extremities on the South Western to North Eastern tips; with a bulky of sampling locations around the central area of the plane (Fig.2.2C and Fig. 2.2D). However, for simplicity of assessments anisotropy was not included in empirical variogram calculations. Therefore, isotropy was assumed in analyses for both sampling schemes.

Figure 2.6 shows the computed empirical variograms (dots) and fitted theoretical variograms (lines) of the spatial coverage sampling and stratified random sampling design. The Gaussian model (Eq. 6) returned the best fit for the stratified simple random sampling scheme while the Spherical model (Eq. 7) provided the best fit for the spatial coverage sampling scheme.

$$\gamma(\mathrm{h}) = \begin{cases} \mathrm{C_0},\ \text{for}\ \mathrm{h} = 0 \\ \mathrm{C_0} + \mathrm{C}\left[1.5\frac{\mathrm{h}}{\mathrm{a}} - 0.5\left(\frac{\mathrm{h}}{\mathrm{a}}\right)^3\right],\ \text{for}\ 0 < h < a \\ \mathrm{C_0} + \mathrm{C},\ \text{for}\ \mathrm{h} \geq \mathrm{a} \end{cases} \qquad (6)$$

$$\gamma(\mathrm{h}) = \mathrm{C}_0 + \mathrm{C}\left[1.5 - \exp\left(-3\left(\frac{-\mathrm{h}^2}{\mathrm{a}}\right)\right)\right],\ \text{for h} \geq 0 \qquad (7)$$

Where $\gamma(\mathrm{h})$ is the semivariance, ‘a’ is the range, ‘h’ is the separation distance, C_0 and C represent the nugget and partial sill, respectively.

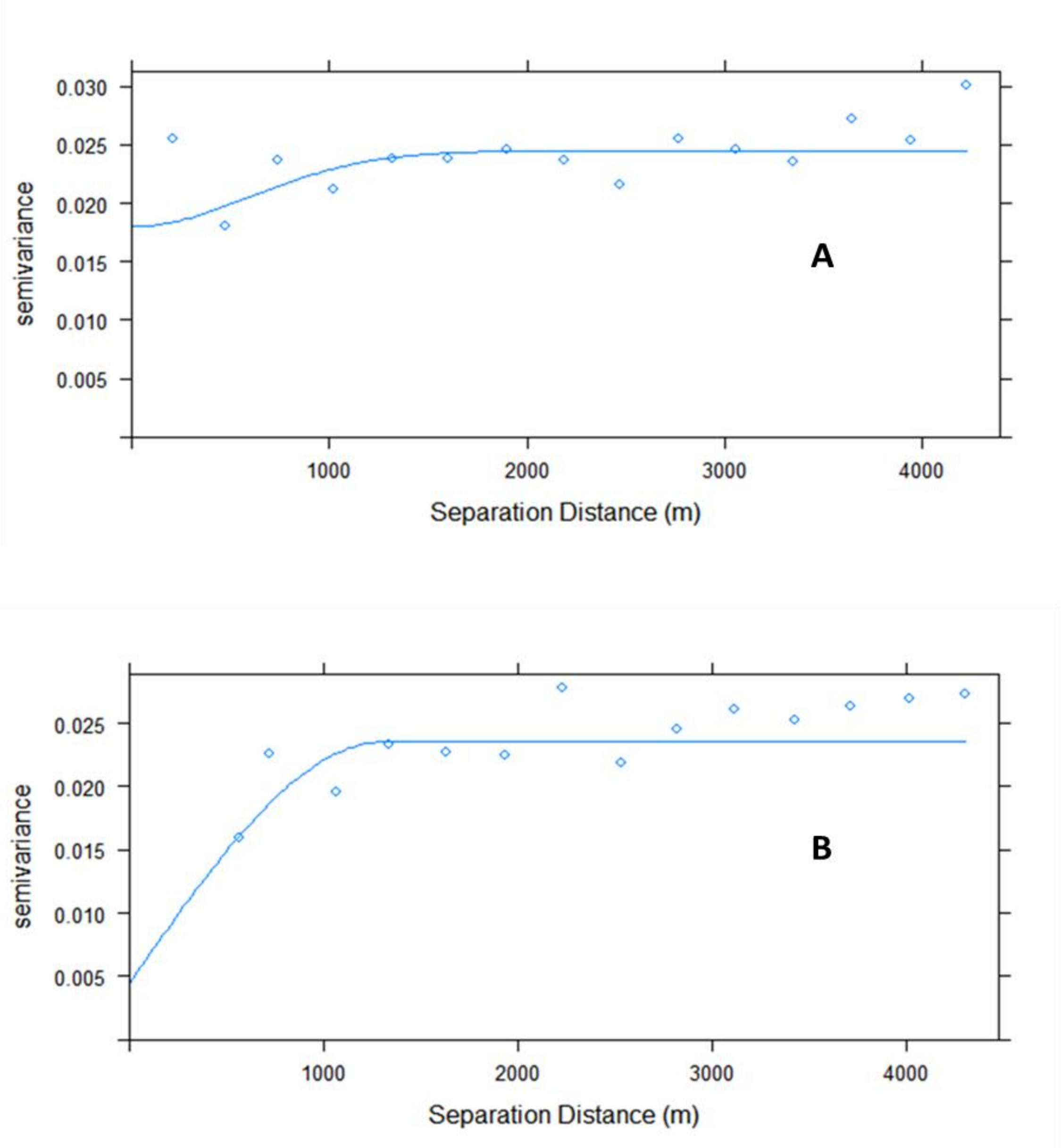

Figure 2.6: Empirical variograms (dots) and fitted theoretical variograms (lines) of STRS approach (A) and SPCS approach (B)

Table 2.2lists the variogram parameters (nugget, partial sill and range) for the respective fitted theoretical variogram models. The range for the STRS scheme was 850 m and 1300 m for the SPCS, respectively. This implies that sampling locations within a separation distance of 850 m for the STRS scheme were spatially correlated. The same was true for locations within the SPCS scheme. The nugget was higher for the STRS scheme indicating a relatively higher variance than the SPCS scheme.

Table 2.2: Variogram Parameter Values

	STRS	**SPCS**
Nugget (C_0)	0.018	0.0045
Partial Sill (C)	0.007	0.019
Range, a (m)	850	1300
Spatial Dependence $\left(\frac{C_0}{C_0+C}\ \mathbf{X\ 100}\right)$	72	19

The spatial dependence ratio for the stratified random sampling scheme (72 %) was higher than that of the spatial coverage sampling scheme (19 %). A variable with a spatial dependence ratio of less than 25 % is considered to have a strong spatial correlation structure, a moderate spatial correlation if the spatial dependence score is between 25 % and 75 %; and a weak spatial correlation if spatial dependence ratio > 75 %, (Sun *et al.*, 2003; Wendroth *et al.*, 2006). The strong spatial correlation structure of the spatial coverage scheme was most likely due to the regular configuration of the sampling nodes leading to a more stable variogram structure. Systematic sampling has been reported to offer superior results especially for regular grid designs (Gao *et al.*, 2012). Spatial coverage sampling is in the category of systematic sampling designs. Sampling nodes in spatial coverage sampling are coincident to the centroids of equal area geostratum (Fig. 2.2C). This configuration leads to a more evenly distributed spread of sampling points across the target area. On the other hand, stratified random sampling sometimes leads to clustering of sampling nodes whereby sampling locations from two adjacent geostrata are positioned on the edge of their respectivegeostratum. Clustering of spatial locations results in a weak spatial structure as evidenced from the spatial dependence ratios of the two approaches (Table 2.2).

Figure 2.7 shows the ordinary kriging map and kriging prediction error surface for the stratified random sampling (STRS) and spatial coverage sampling design (SPCS).

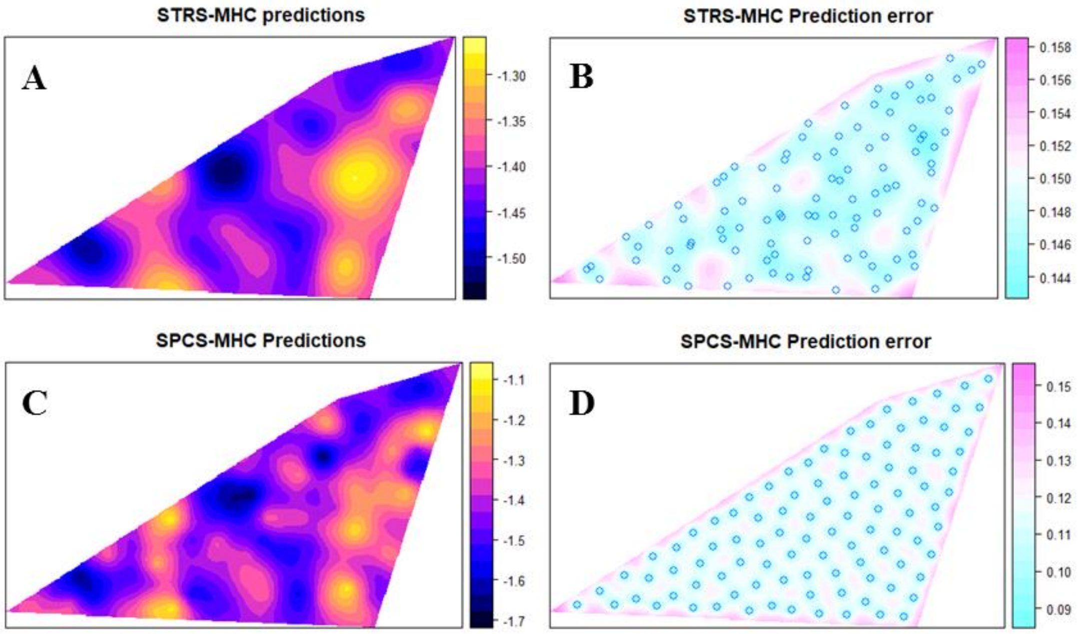

Figure 2.7: Prediction and kriging error surfaces for STRS (A and B) and SPCS Design (C and D)

Predictions for the STRS map ranged between 0.03 and 0.05 cm^{3}cm^{-3} (Fig. 2.7A) while for the SPCS surface ranged between 0.02 and 0.08 cm^{3}cm^{-3} (Fig. 2.7C). Stratified random sampling (STRS) prediction error had lower error values in comparison to the spatial coverage kriging prediction error surface (Fig. 2.7B). The prediction errors for SPCS were uniformly distributed across the surface with the edge-effect is clearly evident on the SPCS prediction error surface (Fig. 2.7D).

Table 2.3 shows the performance indicators of the stratified random sampling scheme and the spatial coverage sampling. The STRS scheme seems to exhibit a slightly better accuracy than the SPCS scheme with smaller ME and RMSE values. The RMSSE score for the SPCS (1.02) indicates a tendency to underestimate MHC.

The STRS design also tends to overestimate MHC. It is evident from the analysis of the performance indicators that the stratified random sampling approach (STRS)results in slightly better estimates for predictions of MHC than spatial coverage sampling (SPCS).

Table 2.3: Performance indicators of the sampling schemes

INDICES	STRS	SPCS
ME	0.0014	0.0017
RMSE	0.156	0.161
RMSSE	0.99	1.02
ASE	0.15	0.14

2.4CONCLUSIONS

In this study a stratified random sampling scheme and spatial coverage sampling scheme were established using a k-means algorithm and compared. Stratified random sampling hada weaker spatial correlation structure than spatial coverage sampling scheme which could be due to a more systematic and stable configuration of sampling locations for the latter approach. The performance indicators for the stratified random sampling scheme (STRS) were slightly better than those for the spatial coverage sampling scheme (SPCS). The study recommends stratified random sampling as an optimal sampling scheme for mapping spatial distributionof soil moisture holding capacity.

2.5REFFERENCES

Asgarzadeh, H., Mosaddeghi M. R., Mahboubi,A. A., Nosrati, A. and Dexter, A. R. (2010). Soil water availability for plants as quantified by conventional available water, least limiting water range and integral water capacity. *Plant Soil* 335:229 – 244.

Beaudette, D.E., Roudier P., and A.T. O'Geen (2012). Algorithms for Quantitative Pedology: A Toolkit for Soil Scientists.[http://r-forge.r-project.org/projects/aqp/] site visited 14/11/2017.

Blake, G.R. and Hartge, K.H. (1986). Bulk density. In: *Methods of Soil Analysis Part 1*:*Physical and Mineralogical Methods. (Edited by Klute, A. et al.)*, Monograph No. 9,Soil Science Society of America, Madison, Wisconsin, USA. pp. 363-375.

Brungard, C.W. and Boettinger, J.L. (2010). Conditioned Latin hypercube sampling: optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In: *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation.*(Edited by Boettinger, J. L., *et al.*). Springer Dordrecht. pp. 67-75.

Brus, D. J. and Heuvelink, G. B. M. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* 138: 86–95.

Brus, D. J., Kempen, B., Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science* 62: 394 – 407.

Brus, D. J. (2015). Balanced sampling: A versatile sampling approach for statistical soil surveys. *Geoderma* 253–254: 111 – 121

Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J. (2013). A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192: 341–353

Chiles, J. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York. 694pp.

De Carlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods* 2(3): 292-307.

Diggle, P. J. and Ribeiro Jr., P. J. (2007). *Model-based Goestatistics*. Springer Science + Business Media, LLC. New York. 228pp.

Doane, D. P. and Seward, L. E. (2011). Measuring Skewness: A Forgotten Statistic? *Journal of Statistics Education* 19 (2): 1-18.

Dumitrescu, A., Birsan, M. V. and Manea, A. (2015). Spatio-temporal interpolation of sub-daily (6 h) precipitation over Romania for the period 1975–2010. *International Journal of Climatology*36 (3): 1331 – 1343.

Gao, Lei, Mingan Shao and Youqi Wang (2012). Spatial scaling of saturated hydraulic conductivity of soils in a small watershed on the Loess Plateau, China. *Journal of Soils Sediments* 12: 863 – 875.

Food and Agriculture Organisation, (2006). *Guidelines for Soil Description*. FAO, Rome, Italy. 66pp.

Gaiser, T., Graef, F. and Cordeiro, J. C. (2000). Water retention characteristics of soils with contrasting clay mineral composition in semi-arid tropical regions. *Australian Journal of Soil Resources* 38: 523 - 536.

Gee, G. W. and Bauder, J. W. (1986). Particle size analysis. In: *Methods of Soil Analysis Part 1: Physical and Mineralogical Methods*. (Edited by Klute, A. et al.), Monograph 9, Soil Science Society of America, Madison, Wisconsin, USA. pp. 383 - 411.

Heuvelink, G. B. M. and Webster, R. (2001). Modelling soil variation: past, present, and future. *Geoderma* 100: 269–301

Heuvelink, G. B. M., Brown, J. D. and Van Loon, E. E. (2007). A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science* 21(5): 497 – 513

Huang, S. W., Jin, J. Y., Yang, L. P., Bai, Y. L. (2006). Spatial variability of soil nutrients and influencing factors in a vegetable production area of Hebei Province in China. *Nutrient Cycling in Agroecosystem*75:201–212.

Li, L., Wu, J., Wilhelm, M., and Ritz, B. (2012). Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in Southern California.*Atmospheric Environment* 1(55): 220–228.

Marques Jr., J., Alleoni, L. R. F., Teixeira, Daniel D. B., Siqueira, D. S., Pereira, G. T. (2015). Sampling planning of micronutrients and aluminium of the soils of São Paulo, Brazil. *Geoderma Regional* 4: 91–99

Minasny, B., Hartemink, A. E., 2011. Predicting soil properties in the tropics. *Earth Science Reviews* 106: 52-62.

Nelson, D. W. and Sommers, L. E. (1982). Total carbon, Organic Carbon, and Organic Matter. In: *Methods of Soil Analysis. Part 2 - Chemical and Mineralogical Properties.* (Edited by Page, A. L. et al.), Monograph 9. American Society of Agronomy, Madison, Wisconsin, USA. pp. 539 - 579.

Novák, V. and Havrila, J. (2006). Method to estimate the critical soil water content of limited availability for plants. Biologia, Bratislava 61(19): 289 - 293,

Odgers, N., McBratney, A. B. Minasny, B. (2011). Bottom-up digital soil mapping. I. Soil layer classes. *Geoderma* 163: 38–44

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* 30: 683-691.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [http://www.R-project.org/]site visited 30/11/2016.

Robinson, T.P. and Metternicht, G. (2006). Testing the performance of spatial interpolation techniquesfor mapping soil properties. *Computers and Electronics in Agriculture* 50: 97–108.

Santra, P., Chopra, U. K. and Chakraborty, D. (2008). Spatial variability of soil properties and its application in predicting surface map of hydraulic parameters in an agricultural farm. *Current Science* 95(7): 937 - 945.

Schoonover, J. E. and Crim, J. F. (2015). An Introduction to Soil Concepts and the Role of Soils in Watershed Management. *Journal of Contemporary Water Research & Education* 154: 21-47

Sun, B., Zhoub, S. and Zhao, Q. (2003). Evaluation of spatial and temporal changes of soil quality based on geostatistical analysis in the hill region of subtropical China. *Geoderma* 115: 85 - 99.

Szatmári, G., Barta, K. and Pásztor, L. (2015). An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping. *Hungarian Geographical Bulletin* 64 (1): 35–48.

Walvoort, D.J.J., Brus, D.J., de Gruijter, J. J. (2010). An R-package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences* 36: 1261–1267

Webster, R. and Lark, M. (2013). *Field Sampling for Environmental Science and Management*. Routledge, Oxford. 192pp.

Wendroth, O., Koszinski, S. and Pena-Yewtukhiv, E. (2006). Spatial association among soil hydraulic properties, soil texture, and geoelectrical resistivity. *Vadose Zone Journal* 5: 341 - 355.

CHAPTER THREE

3.0PREDICITION OF SOIL MOISTURE HOLDING CAPACITY WITH SUPPORT VECTOR MACHINES IN DRY SUB-HUMID TROPICS

ABSTRACT

Adaptation of the agro-hydrological function of dry sub-humid environments to changing moisture regimes is important for sustaining crop yields. Soil moisture holding capacity data is required for modelling to this end but is hardly sufficient and costly tomeasurein the field. An alternative approach is the use of mathematical equations called pedotransfer functions (PTFs) which use readily available soil physicochemical properties as inputs to estimate soil moisture holding capacity. Several regression techniques have been established and applied for developing PTFs with some shortcomings. This study explored the application of a promising data mining method known as support vector machines (SVM) in the development of PTFs for dry sub-humid tropical soils. A soil dataset consisting of 296 samples of measured moisture content and soil physicochemicalproperties wasused. Support vector machines PTFs models to estimate moisture content were developed in R-software. Model predicted moisture content was compared against measured moisture values based on the Root Mean Square Error and Mean Error and coefficient of determination (R^2) as performance indices. Developed support vector machines PTFs had better accuracy than published SVM-PTFs and can be integrated in a modelling framework for estimation of soil moisture holding capacity.

Keywords: moisture holding capacity, pedotransfer functions, support vector machines, sub-humid tropics

3.1INTRODUCTION

Dry sub-humid zones in Tanzania are regions of marginal agricultural productivity (Kilasara, 2010) and highly sensitive to seasonal moisture availability. Long term sustainability of crop yields will require integrated modelling approaches to provide the necessary feedback for adapting agrohydrological functions to changing soil moisture regimes (Vereecken *et al.*, 2016). Soil moisture holding capacity is an important parameter for modelling agricultural productivity of sub-humid zones. It is a measure of the difference between moisture at field capacity and wilting point. Moisture holding capacity facilitates the description of soil hydrological processes such as drainage, infiltration and percolation and is vital input data in models such as Soil Water Assessment Tool (SWAT) (Toth *et al.*, 2015), and AQUACROP (Raes *et al.*, 2009).

Model results are highly dependent on the nature and quantity of data (Wosten *et al.*, 2013; Toth *et al.*, 2015), but soil moisture data is generally in limited supply for tropical soils (Schaap, 2005; Minasny and Hartemink, 2011) largely due to high costs of measurement and lack of associated equipment. Mathematical equations known as pedotransfer functions (PTFs), linking easily measured soil properties as input variables to soil moisture data, have been employed to bridge data gaps. With extensive development for temperate soils (Nguyen *et al.*, 2014), PTFs application is fraught with specificity to calibration datasets (Vereecken *et al.*, 2016) and geographic regions (Haghverdi *et al.*, 2012; Nguyen *et al.*, 2014). Tropical soils have a bimodal particle size distribution in contrast to the uni-modal soils of the temperates (Condappa *et al.*, 2008; Minasny and Hartemink, 2011), with maximal weight percentage for clay and sand size fractions and low silt content (Minasny and Hartemink, 2011). This is suggested to impart contrasting soil hydraulic characteristics (Wosten *et al.*, 2001; Minasny and Hartemink, 2011; Vereecken *et al.*, 2010; Botula *et al.*, 2013), limiting transferability of PTFs for modelling processes across their statistical and pedo-climatic calibration bounds (Haghverdi *et al.*, 2012).

Utility of PTFs necessitates validation or development of new PTFs for improved modelling outputs (Vereecken *et al.*, 2010; Haghverdi *et al.*, 2012). Studies to this end for tropical soils in sub-Saharan Africa include Young *et al.* (1999), Mdemu and Mlengera (2002), Mugabe (2004), Obalum and Obi (2013), Botula *et al.* (2012; 2013), Wosten *et al.* (2013), andMdemu (2015). All these studies have shortcomings including evaluation on small soil datasets or compiled soil databases or frequent application of multiple linear regression method. Among the many PTF-development methods, the multiple linear regression method has also been highlighted to be inflexible in capturing the non-linearity associated with moisture holding properties (ReF).An insufficient data size has been reported to be a major drawback for PTF evaluations (Nemes *et al.*, 2006; Pachepsky and Rawls, 1999). Substantial uncertainty also exists with soil databases used to derive the PTFs (Vereecken *et al.*, 2010), probably associated with data entry or measurement inconsistences (Vereecken *et al.*, 2010; Minasny and Hartemink, 2011).

Machine learning algorithms generally have better flexibility in mimicking the complex nonlinear pattern in the soil-moisture continuum (Botula *et al.*, 2013).

Ubiquity of computer technology and enhanced computational efficiency has spiralled the advancement of sophisticated machine learning algorithms such as artificial neural networks (Agyare *et al.*, 2007; Haghverdi *et al.*, 2012), k-Nearest neighbour (Nemes *et al.*, 2006; Botula *et al.*, 2013; Nguyen *et al.*, 2015), and Support Vector Machines (SVMs) (Twarakavi *et al.* 2009; Kovačević *et al.*, 2010; Lamorski *et al.*, 2014; Khlosi *et al.*, 2016). The respective authors provide a good technical background about these highlighted methods. Interest here is skewed to the SVMs because they have circumvented typical drawbacks associated with the popular ANNs (Balabina and Lomakina, 2011; Haghverdi *et al.*, 2014).

Support vector machines are a supervised machine learning algorithm based on statistical learning theory (Wu *et al*., 2003; Li *et al*., 2014). Support vector machines were developed by Vapnik (1995) for data classification and later extended to solve regression problems (Wu *et al*., 2004; Balabina and Lomakina, 2011; Haghverdi *et al*., 2014). Lamorski *et al.* (2008) and Twarakavi *et al.* (2009) pioneered the use of SVMs in the development of PTFs for the parametric functions and soil matric points, reporting improvements. The key advantage of the SVMs is structural risk minimisation over the empirical risk minimisation which checks overfitting during model development (Twarakavi *et al*., 2009; Wu *et al*., 2003). The SVM technique is also easier to implement than ANNs (Hsu *et al.*, 2016). Interest in the use of SVMs for PTF development has been stimulated (Haghverdi *et al*., 2014; Nguyen *et al*., 2015; Khlosi *et al*., 2016) but with no work evident for sub-Saharan Africa soils. Flexibility of SVMs in incorporating new soil data would be of added benefit particularly for developing countries, where soil datasets are in high demand for simulating the agro-hydrological function and productivity of farming systems. In view of this, the objective of this research was to apply support vector machines to develop pedotransfer functions for moisture holding capacity using experimentally measured data. Henceforth, the expression support vector regression (SVR) is used instead of support vector machines (SVMs) to stress its application to regression other than classification.

3.2MATERIALS AND METHODS

3.2.1 Studyarea

The study area was Ilakala village inKilosa District, Morogoro Region,Tanzania (Fig. 3.1);within latitudes 7^{o} 5’ 30” S and 7^{o} 9’ 30” S and longitudes 36^{o} 50’ 30” E and 36^{o} 57’ 30” E. It has a total area of about 44 km^{2}. Agriculture (both cropping and livestock keeping) is the major livelihood activity in the area. The cropping system is a maize-sesame-pigeon peas small-holder system; with maize and pigeon peas as the main food crops. Sesame is a cash crop. Livestock keepingis typically undertaken by pastoralist communities of Masai and Sukuma ethnicities.

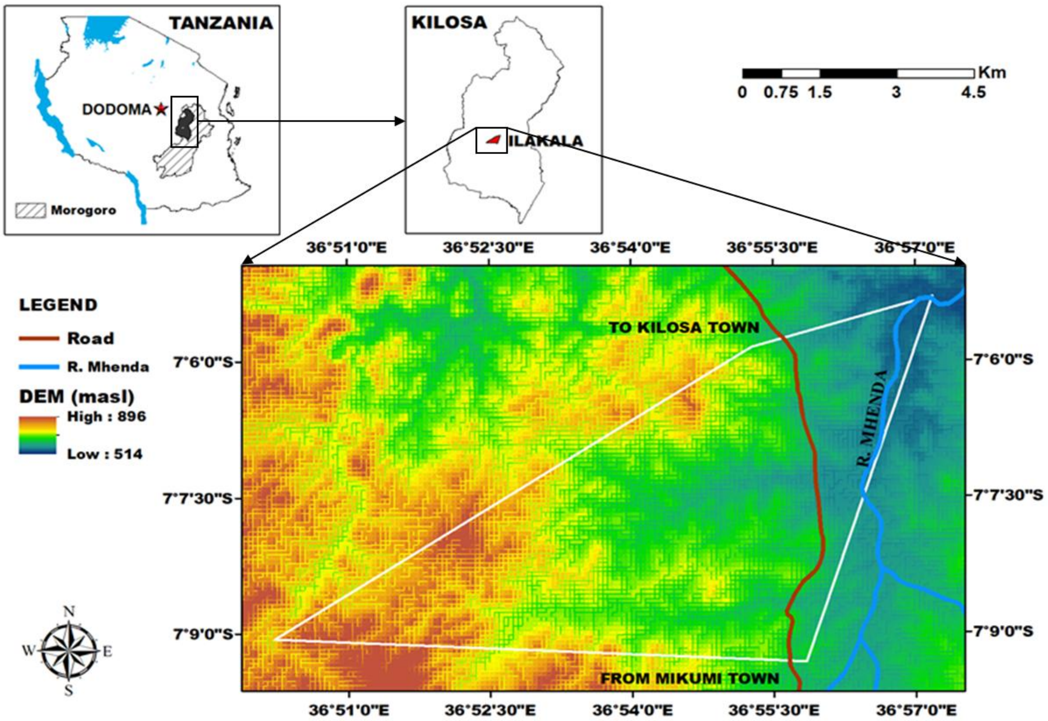

Figure 3.1: Location map of Ilakala Village

The area is mostly traversed with soils of a sandy textural origin. Major soil types are Hyperdystric Cambisol (loamic, ochric), Rhodic Acrisol (clayic, cutanic, epieutric, profondic), Luvic Stagnic Umbrisol (endoeutric, loamic), Endogleyic Protovertic Eutric Cambisol (colluvic, ruptic), and Pellic Vertisol (ferric, humic, mesotrophic) (Kaingo and Tumbo, 2016). Many seasonal streams drain the area from the hilly regions in the southwest and western, feeding into River Mhenda that flows along the eastern edge of the village.

3.2.2 Soil dataset

A soil dataset of 296 samplescollected between June 2014 and July 2015was used in this study. Soil samples were taken from 100 spatial locations at three depths (0-30 cm, 30-60 cm and 60-100 cm). However, soil samples at the 60-100 cm depth interval were not taken at four sampling locations due to rockiness. The sampling scheme and laboratory analyses of the soil samples are described in section 2.2.2 and 2.2.3, respectively (Chapter Two of this dissertation).Constitutive soil variables were bulk density (BD), soil organic carbon (OC), sand, clay and silt content, and moisture content at field capacity (FC)and wilting point (WP). Soil matric suctions of 30 kPa and 1500 kPa were used as the FC and WP points, respectively. The dataset was randomly split into a ratio of 2:1 fora training dataset (n=198) and testing set (n=98), respectively. Descriptive statistics, gaussian tests and correlation analyses were performed using R-software (R Core Team, 2016).

3.2.3 PTFs development

The training dataset was used for the SVR model calibration. Epsilon support vector regression (ε-SVR) was used for development of SVR PTFs in the e1071 R-software package. The success of the SVR calibration process depends on three key issues: the choice of the cost/regularisation parameter - C, the "tube" insensitivity variable (ε) and the priori selection of kernel function (Twarakavi *et al.*, 2009). Table 3.1 shows the common kernel functions for SVR. The radial basis function (RBF) kernel is most frequently used but a linear kernel was chosen for this study because of the overfitting challenges associated with the RBF kernel (Lamorski *et al.*, 2014).Overfitting occurs when a model starts to describe the random error in the data rather than the relationships between variables. An overfit model reduces its generalisability outside the original dataset (Babyak, 2004).

Table 3.1: Common kernel functions and their hyper-parameters in SVR

Kernels	Functions	Parameters
Radial basis function	$\boldsymbol{e}^{(-\gamma\|x_i^T - x_j\|^2)}$	C, ε, γ
Linear	$\boldsymbol{x_i^T . x_j}$	C, ε
Polynomial	$(\boldsymbol{\gamma x_i^T . x_j + r})^{\boldsymbol{d}}$	C, ε, r, d

The C-parameter determines tolerance of the calibration prediction error and the structural complexity of the SVR model. A large C value results in model complexity leading to a computationally inefficient model, overfitting and poor generalisation capability. The ε-parameter controls the loss function which controls the width of the insensitive zone leading to minimising of the regression risk. Large values of ε lead to smaller numbers of support vectors and poor generalisation. Parameters C and ε are known as the hyper-parameters and their optimisation determines how good the SVR model is. While 'γ', 'r', and 'd' are kernel parameters.

The SVR calibration procedure was carried out in three steps. First, the training dataset was used to initially fit the SVR model with the linear kernel function through epsilon-regression in e1071 R-software package. Linear kernel functions have only two hyper-parameter values that require setting i.e. the C and ε-parameter. The default package C-parameter value (C=1) was retained and the ε-parameter set to 3σ√(ln(n)/n) following Ließ *et al.* (2016) for the initial fit, where n is the number of records in the training dataset and σ the standard deviation of the data. In the second step, tuning of the SVR model hyper-parameters was performed using a grid-search method with a 10-fold cross validation in 5 repeats. The grid-search method facilitates optimization ofhyper-parameters by estimating training prediction error for each set of all possible combinations of hyper-parameters within the feasible feature space (Twarakavi *et al.*, 2009). With insights from earlier studies (Lamorski *et al.*, 2008; Haghverdi *et al.*, 2014), parameter search space was a priori set to; $0.001 \leq C \leq 1000$ at an incremental ratio of 10, $0 \leq \varepsilon \leq 0.3$ at steps of 0.001. Subsequent fine tuning was performed using a parameter search space within the neighborhood of the best optimized hyper-parameters from the second step. This process generated the best optimal hyper-parameters that were used for ultimately developing the SVR – PTFs in the third step. For comparison purposes, multiple linear regression (MLR) based PTFs were also developed. Step-wise regression was used to develop the MLR-PTFs using the SPSS software package version20 (IBM Corp., 2011). Both the SVR-PTFs and MLR-PTFs were then applied to the testing data to assess their validity. Performance of the developed PTFs was evaluated using the Root Mean Square Error (RMSE), Mean Error (ME) and coefficient of determination (R^2) as indicators. The RMSE and ME should ideally be close to zero while the R^2 should be closer to one.

$$\mathbf{ME} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}[\mathbf{y} - \hat{\mathbf{y}}\,]$$

$$\mathbf{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}[\,\mathbf{y} - \hat{\mathbf{y}}\,]^{\mathbf{2}}}$$

$$\mathbf{R} = \mathbf{1} - \frac{\sum_{i=1}^{n}[\,\mathbf{y} - \hat{\mathbf{y}}\,]^2}{\sum_{i=1}^{n}[\,\mathbf{y} - \bar{\mathbf{y}}\,]^2}$$

Where y is the measured moisture content, ŷ is the predicted moisture content,ȳ is the mean of the measured moisture content, and ‘n’ the number of datasets.

3.3RESULTS AND DISCUSSION

3.3.1 Descriptive statistics of soil datasets

Table 3.2 shows the summary statistics of the training and testing datasets. Across both datasets bulk density ranged from 1 to 1.19 g/cc. Organic carbon ranged from 0.06 %. to 3.37 %. Clay, sand, and silt content ranged from 0.1 % to 63.6 %, 20 % to 96.6 %, and 1.4 % to 35.4 %, respectively. Moisture content at FC (θ_{30}) ranged from 0.08 to 0.48 cm^3cm^{-3} while moisture at WP (θ_{1500}) ranged from 0.03 to 0.39 cm^3cm^{-3}. Mean values of training and testing datasets were similar for all soil variables. Though the skewness indices were consistent with a Gaussian symmetrical distribution (Doane and Seward, 2011), Kurtosis values were non-optimal for an assumption of normality to be held (DeCarlo, 1997).

Table 3.2: Descriptive statistics of Training and Testing datasets

	Min	Max	Mean	SD	Skewness	Kurtosis
	Training					
BD	1.00	1.19	1.06	0.04	1.17	1.40
OC	0.06	3.23	0.80	0.58	1.63	3.30
CLAY	0.10	63.60	22.19	16.64	0.64	-0.53
SAND	20.00	96.60	64.90	16.41	-0.52	-0.35
SILT	1.40	35.40	12.90	4.89	0.87	2.75
θ_{30}	0.08	0.48	0.23	0.07	0.52	0.23
θ_{1500}	0.03	0.38	0.18	0.07	0.31	-0.13
	Testing					
BD	1.00	1.16	1.05	0.03	0.97	0.88
OC	0.10	3.37	0.82	0.64	1.69	3.42
CLAY	0.10	61.00	22.26	15.78	0.55	-0.64
SAND	25.60	94.60	64.67	15.97	-0.31	-0.61
SILT	2.80	24.40	13.08	5.29	0.10	-0.71
θ_{30}	0.09	0.41	0.23	0.07	0.02	-0.46
θ_{1500}	0.05	0.39	0.19	0.07	0.05	-0.26

Table 3.3 shows the correlation coefficients for the soil physicochemical properties on moisture content at FC (θ_{30}) and moisture content at WP (θ_{1500}). Sand and Clay had a strong correlation ($r > 0.7$) but with opposite polarity for both FC and WP. Bulk density, OC, clay and sand had highly significant correlations with moisture content at θ_{30} and θ_{1500}. Silt was poorly correlated to θ_{30} and θ_{1500} with $r < 0.07$. Organic carbon was positively correlated with moisture content at both suction extremes. Organic carbon content influences moisture retention properties due to its role on many other physical and physico-chemical soil properties. Higher OC content improves soil structure and porosity, leading to increased moisture-holding

capacity (Toth *et al.*, 2015). Organic matter also has high cation exchange capacity and high specific surface area which enhances its moisture-absorption properties (Kaingo, 2011).

Table 3.3: Pearson correlation coefficients for soil variables

	BD.g.cc.	OC	CLAY	SAND	SILT
θ_{30}	-0.46***	0.23***	0.73***	-0.76***	0.07
θ_{1500}	-0.46***	0.28***	0.77***	-0.8***	0.07

3.3.2 PTFs Development

Input θ_{30} and θ_{1500} data were log-transformed prior to model development for both the MLR and SVR method. This was necessary for consistency of the target variables with the normal probability distribution. The initial fit generated SVR-models with support vectors ranging from 188 to 191 at hyper-parameter settings of C=1 and $\varepsilon = 0.034$ derived from$3\sigma\sqrt{}(\ln(n)/n$ (results not shown). This translated to about 95 % -96 % of the total support vectors used in model formulation, suggesting poor generalisation of the models with this initial choice of hyper-parameters. The number of support vectors within the SVR model signifies its suitability for predictions on a new dataset. A larger proportion of support vectors lead to overfitting of the model and poor predictions on a new dataset, while a smaller proportion leads to under-fitting (Twarakavi *et al.*, 2009). A 50 % threshold has been held as the theoretically optimal proportion of support vectors for good generalisation on new datasets (Twarakavi *et al.*, 2009; Lamorski *et al.*, 2014).

Figure 3.2A to 3.2H show model sensitivity with varying SVR hyper-parameters combinations andincreasing soil predictor variables during the coarse grid search tuning process. Cross-validation error for FC SVR-models ranged between 0.035 to 0.08 (Fig. 3.2A – Fig. 3.2D), respectively, corresponding to model types FC1 to FC4 (Table 7), while for WP SVR-models ranged between 0.04 to 0.14 ((Fig. 3.2E – Fig. 3.2H) for model types WP1 to WP4.Models were most sensitive to values of C-parameters. Generally, with lower C values ($C<10^{-2}$)leading to higher errors for both WP and FC SVR-models. The WP models were most sensitive to the hyper-parameters than the FC model withthe CV errors for the WP models almost twice those of FC model for the same predictor variables(Fig. 3.2A vs Fig. 3.2E, Fig. 3.2B vs Fig. 3.2F, Fig. 3.2C vs Fig. 3.2G). However, that was not the case for FC4 and WP4, which showed similar CV errors because the WP4 had an extra predictor variable (OC) than FC4.

The pattern of higher CV errors for the WP SVR-models was perhaps related to the input predictor variables in the model. Inclusion of predictors in a model was arbitrary guided by the correlation results (Table 3.3). Much as sand was highly correlated toθ_{1500}, its inclusion as a predictor in WP modelsis suspect. At high soil matric suctions (i.e. 1500), moisture in the soil matrix is greatly influenced by the specific surface area and capillary forces (Khorshidi, 2015). Sand has plenty of macrovoids which hold moisture most at the low matric suctions and thus have less influence at high suctions because of its lower cation exchanges capacity (Khorshidi, 2015).

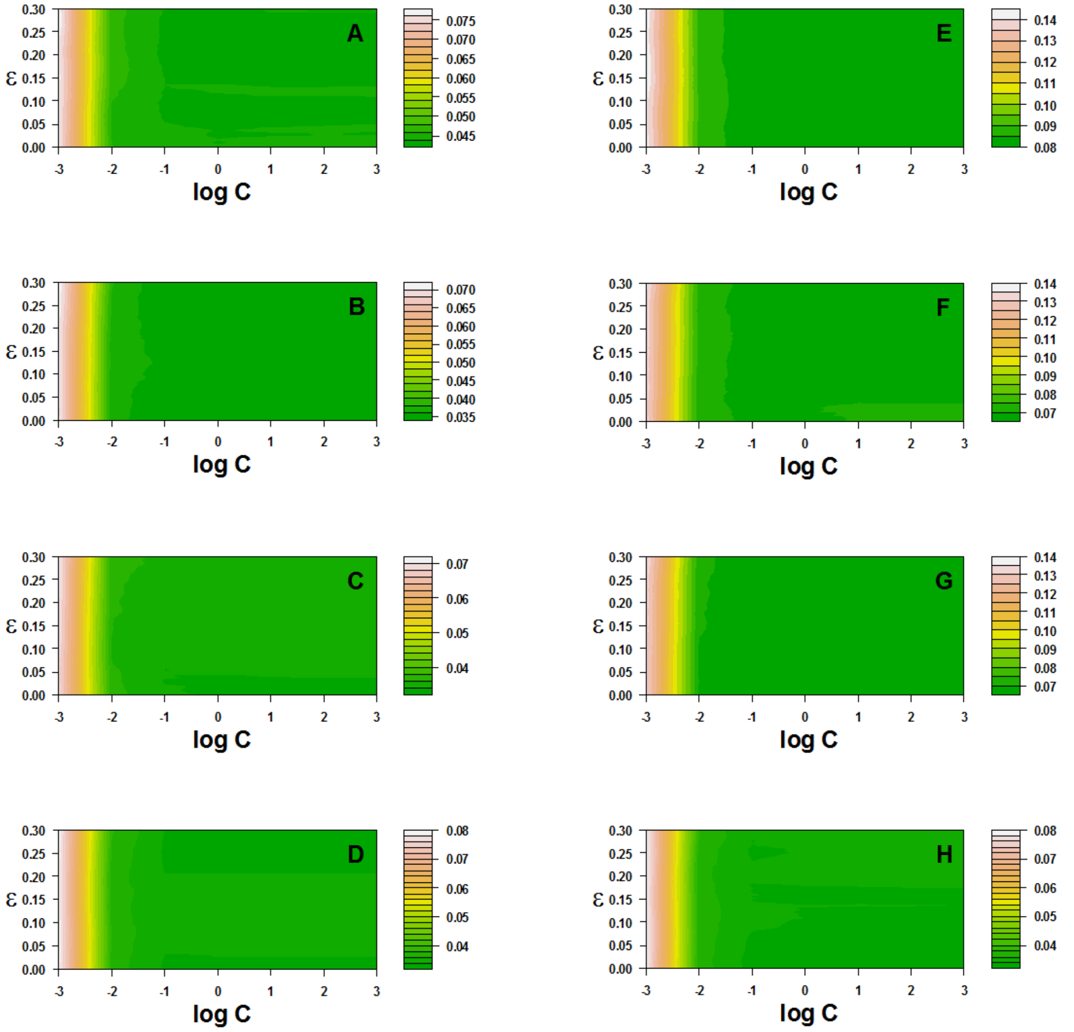

Figure 3.2: **Sensitivity of SVR hyper-parameter calibration to incremental soil predictor variables (A=FC1, B=FC2,C=FC3,D=FC4,E=WP1, F=WP2,G=WP3,H=WP4)**

Table 3.4 shows the most optimal hyper-parameters from the coarse grid-search (1st) and fine grid search (2nd) processes, with their corresponding cross-validation errors (CVerror) and number of support vectors (SVs). A lower number of SVs were evident for the SVR models after the 2nd tuning except for the FC4 model.

Table 3.4: Calibration results of optimal hyper-parameters of SVR-model types

	PREDICTORS	C		ε		SVs		CVerror	
		1st	2nd	1st	2nd	1st	2nd	1st	2nd
FC1	SAND+CLAY	100	127	0.197	0.21	151	143	0.043	0.044
FC2	CLAY+SAND+BD	1	0.1	0	0.25	198	128	0.034	0.034
FC3	CLAY+SAND+BD+OC	100	96	0.006	0.022	196	189	0.034	0.034
FC4	SAND+BD	1	0.4	0.244	0.24	126	128	0.034	0.034
WP1	SAND+CLAY	0.1	0.35	0.199	0.217	146	138	0.080	0.081
WP2	CLAY+SAND+BD	1000	950	0.246	0.24	119	124	0.067	0.067
WP3	CLAY+SAND+BD+OC	0.1	0.65	0.15	0.15	151	149	0.065	0.065
WP4	SAND+BD+OC	10	28	0.037	0.156	183	150	0.033	0.065

Table 3.5 shows the coefficients for the MLRs-PTFs. All input variables were statistically significant. The tolerance and VIF scores indicate that the soil variables included as inputs were important predictors for moisture retention at FC and WP. A tolerance score > 0.1 and VIF < 10 indicate absence of multicollinearity and hence model parsimony.This implies that only the most influential predictor variables were objectively retained in the regression model.Bulk density has an inverse influence on the prediction of moisture retention. Increase in bulk density results in the destruction of pedostructure and pore architecture leading to a reduction in the available volume for soil moisture storage. Sand as a predictor variable had an inverse and the least influence on the moisture predictands in the MLR model. This trend could be explained by increases in soil macropores associated with sandy soils which causes a decline in moisture retention(Tuller and Or, 2001). Further, sand particle fractions have a low cation exchange capacity (IPNI, 2011) which results in limited adsorptive sites for retaining moisture (Khorshidi, 2015). The small β-

coefficient for OC could have been because of the calibration of the model on a dataset with a low OC content. Average value for OC content in the training dataset was 0.8 % (Table 3.2), which corresponds to a classification class of very low.

Table 3.5: β-Coefficients of MLR PTFs for FC and WP

Variable	β	Sig.	Tolerance	*VIF
	FC			
Intercept	2.163	0		
Sand	-0.013	0	0.926	1.08
Bulk Density	-2.712	0	0.926	1.08
	WP			
Intercept	2.916	0		
Sand	-0.017	0	0.855	1.17
Bulk Density	-3.493	0	0.911	1.097
OC	0.083	0.011	0.921	1.086

*VIF – Variance Inflation Factor; Sig- Significance at 5 % probability level.

3.3.3 Evaluation of PTFs

Table 3.6shows the performance indicators for the SVR models with varying predictors. The RMSE values ranged from 0.037 cm^3 cm^{-3} to 0.042 cm^3 cm^{-3}. The MEs for the developed SVR models except model FC3 were less than zero indicating a tendency to underestimate moisture at FC and WP. Coefficients of determination (R^2) were between 56.2 % to 67.9 % but slightly higher for the SVR models for wilting point (WP2, WP3, WP4) than the field capacity models (FC2, FC3, FC4). Best SVR models were FC3 for moisture prediction at field capacity with sand, clay, bulk density and organic carbon as predictors. For wilting point, model WP4 was the best performing model with sand, bulk density and organic carbon as predictors. The developed models explain a substantial proportion of variance of the data and provide satisfactory quantitative estimates of moisture. According to Gholizadeh *et al.* (2015), models with R^2 values of 0.50 to 0.65 show good discrimination between low and high values while those within 0.66–0.81 indicate approximate quantitative predictions, 0.82– 0.90 good prediction, with R^2> 0.91 are excellent.

Table 3.6: Performance Indicators for SVR models with different Predictors

MODEL	INPUTS	ME	RMSE	R^2
FC1	SAND+CLAY	-0.006	0.042	0.562
FC2	CLAY+SAND+BD	-0.003	0.038	0.643
FC3	CLAY+SAND+BD+OC	0.000	0.037	0.663
FC4	SAND+BD	-0.003	0.038	0.645
WP1	SAND+CLAY	-0.007	0.045	0.546
WP2	CLAY+SAND+BD	-0.004	0.037	0.668
WP3	CLAY+SAND+BD+OC	-0.003	0.037	0.677
WP4	SAND+BD+OC	-0.003	0.037	0.679

Unit plots of SVR and MLR predicted moisture content on the testing dataset are shown in Fig. 3.3. The best performing SVR PTFs (SVR-FC3 and SVR-WP4) were compared here. The R^2, ME and RMSE values were marginally better for the SVR PTFs (upper panel) than the developed MLR-PTFs (lower panel). Prediction indices are better at wilting point than at field capacity for both SVR and MLR models. Miháliková *et al.* (2016) also found moisture content predictions to be more reliable at WP than at FC.

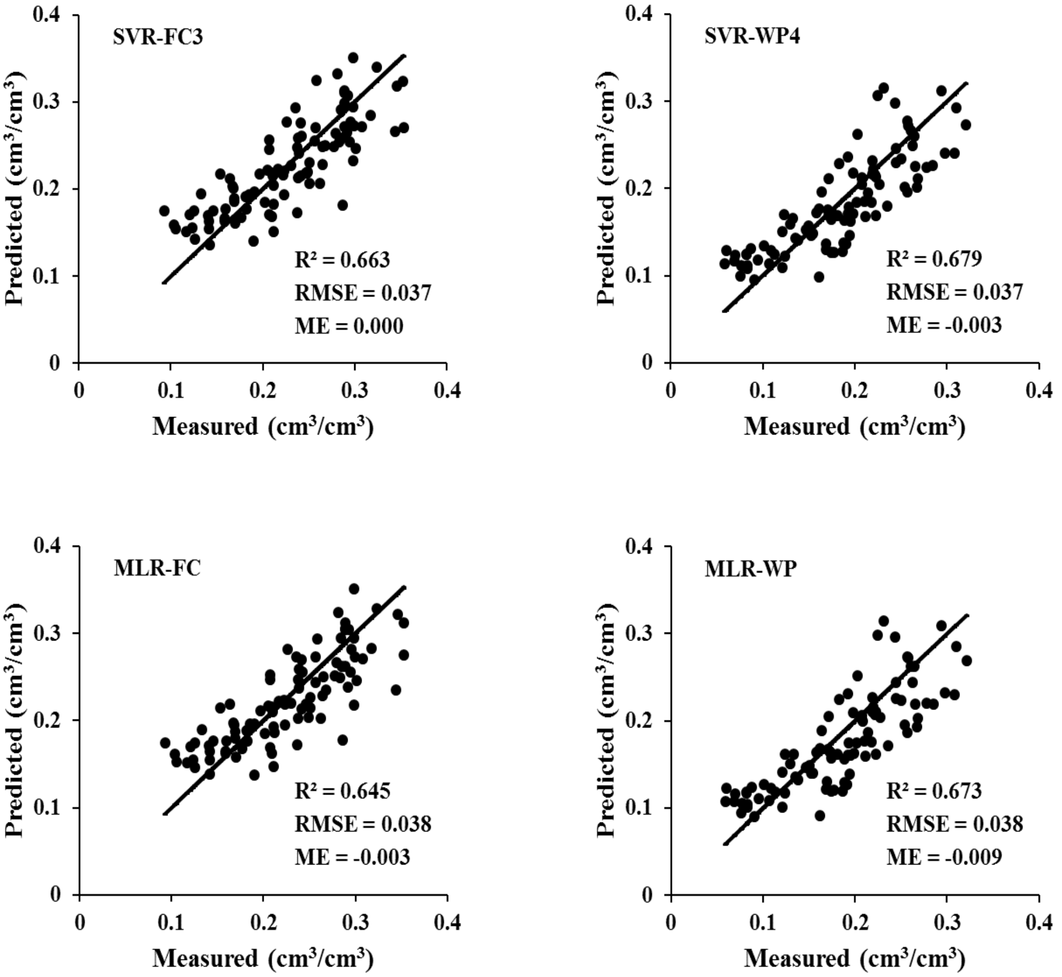

Figure 3.3: Unit plots of SVR and MLR predicted moisture content on the testing dataset

The possible reason for this trend could be linked to the fact that moisture content at higher matric potentials (FC) is controlled by numerous soil factors which results in large variability within measurements. In contrast, moisture at wilting point is mainly influenced by specific surface area of the soil constituents which minimises variability in measurement values. The negative ME values indicate that models tend to underestimate moisture content at FC and WP for both the MLR and SVR-PTFs. Though the ME of SVR-FC3 (ME=0.000) might suggest an unbiased model, this result was due to the deviations above and below the line of fit cancelling out.

The RMSE values for both the FC and WP were 0.037 cm^3 cm^{-3}, for the MLR-PTFs and 0.038 cm^3 cm^{-3} for the SVR models. These RMSE values are within typical RMSE values for PTFs reported to range within 0.02 and 0.07 (Rawls and Pachepsky, 1999), suggesting a good model accuracy. The RMSE values of these SVR-PTFs were lower than those reported in similar works (Lamorski *et al.*, 2008; Twarakavi *et al.*, 2009; Lamorski *et al.*, 2014; Nguyen *et al.*, 2015; Khlosi *et al.*, 2016), for matric potentials at or near FC or WP. The explanation for this is not straight forward and can only be suggested. The trend could be associated with the choice of kernel function adopted in the SVR model development. Haghverdi *et al.* (2014) noted that low RMSE values were associated with models that predict linear matric potential-moisture content relationship than the nonlinear relationship. This is a plausible link as a linear kernel was adopted in this study while the radial basis function (RBF) kernel was used in those studies.

In comparing the linear and RBF kernels, Lamorski *et al.* (2014) found that using the RBF kernel in development of SVR models led to overfitting with high RMSE and poor generalisation capability of the models on validation. He concluded that the pattern was related to the high sensitivity of the SVR-models to the γ-parameter of the RBF kernel. Another possible explanation could be the variations in dataset characteristics used in the different studies as well as the predictors adopted for the SVR. Differences in measurement approaches and textural composition of the samples in datasets induce variability which affects the quality of PTF outputs (Vereecken *et al.*, 2010). Including additional predictors to the particle size fractions improved accuracy of the SVR-models (Twarakavi *et al.*, 2009; Nguyen *et al.*, 2015;

Khlosi *et al.*, 2016). A similar trend was observed in this study; however, careful thought is needed to avoid including difficult to measure soil properties as predictors as this contradicts the essence of PTFs.

Complete focus on the performance indices (Table 3.6) may mask overfitting issues that might arise when additional predictors are included in the SVR-models. Though the SVR-FC3 model with sand, clay, BD and OC as predictors (Fig. 3.3) has better indices than the SVR-FC4 (Fig. 3.4) with only sand and bulk density as predictors, its proportion of support vectors (SVs=189, Table 3.4) was higher than those of the SVR-FC4 model (SVs= 128). This has implications for the generalisation capability of the model as earlier alluded to. Reliability of the SVR-FC4 model would be better when applied to independent datasets than the SVR-FC3 model. It is important to note that the difference in the indices is only marginal between the two model types.

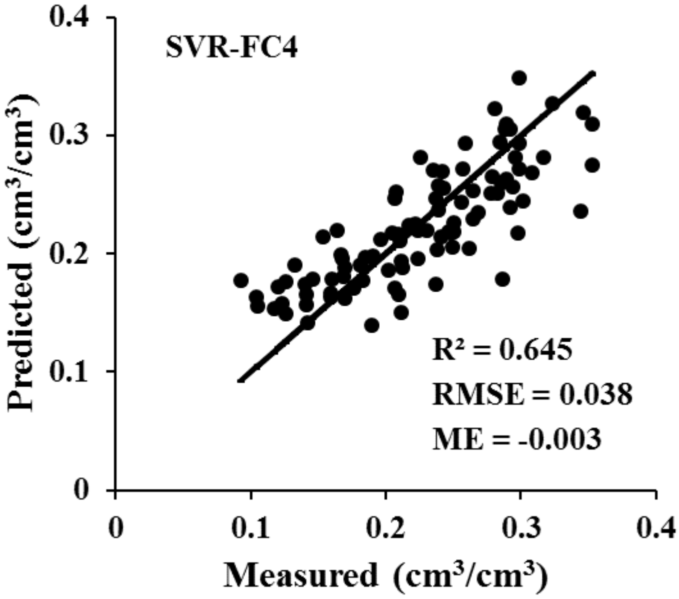

Figure 3.4: Unit plots of SVR-FC4 PTF on the testing dataset

3.5CONCLUSIONS

This study was undertaken to develop SVR pedo-transfer functions for estimating soil moisture holding capacity for dry sub-humid soils. The SVR-PTFs developed in this study performed slightly better than published SVR PTFs. The developed multiple linear regression-PTFs are a suitable straight forward application as an alternative.

3.6REFERENCES

Agyare, W. A., Park, S. J. and Vlek, P. L. G. (2007). Artificial neural network estimation of saturated hydraulic conductivity. *Vadose Zone Journal* 6: 423 – 431.

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models.

Balabina, R. M. and Lomakina, E. I. (2011). Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. *Analyst* 136:1703–1712.

Botula, Y. D., Nemes, A., Mafuka, P., Van Ranst, E. Cornelis, W. M. (2013). Prediction of Water Retention of Soils from the Humid Tropics by the Nonparametric -Nearest Neighbor Approach. *Vadose Zone Journal*12 (2): 1-17

Condappa, D., Galle, S., Dewandel, B. and Haverkamp, R. (2008). Bimodal zone of the soil textural triangle: Common in Tropical and Subtropical Regions. Soil *Science Society of America Journal* 72: 33 - 40.

De Carlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods* 2(3): 292-307.

Doane, D. P. and Seward, L. E. (2011). Measuring Skewness: A Forgotten Statistic? *Journal of Statistics Education* 19 (2): 1 – 18.

Gholizadeh, A., Borůvka, L., Saberioon, M. M., Kozák, J., Vašát, R. and Němeček, K. (2015). Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. *Soil and Water Research10* (4): 218 – 227.

Haghverdi, A., Cornelis, W. M., Ghahraman, B. (2012). A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. *Journal of Hydrology* 442–443: 46–54.

Haghverdi, A., Öztürk, H. S., Cornelis, W. M. (2014). Revisiting the pseudo continuous pedotransfer function concept: Impact of data quality and data mining method. *Geoderma* 226–227: 31–38

Hsu, C. W., Chang, C. C., and Lin, C. J. (2016). A Practical Guide to Support Vector Classification.[http://www.datascienceassn.org/sites/default/files/Practical%20Guide%20to%20Support%20Vector%20Classification.pdf] site visited on 31/10/2017.

International Plant Nutrition Institute (2011). *Cation Exchange: A Review.* Insights. 4pp

Kaingo, J. (2011). Spatial prediction of soil water retention in the Ngerengere sub-catchment, Morogoro-Tanzania. Dissertation submitted for Award of MSc degree at Sokoine University of Agriculture, Tanzania. 110pp.

Kaingo, J. and Tumbo, S. D. (2016). Soil Mapping and Web-GIS Development for Trans-Sec Project: Final Report. Sokoine Univesity of Agriculture, Morogoro - Tanzania. 88pp.

Khlosi, M., Alhamdoosh, M., Douaik, A. Gabriels, D. and Cornelis, W. M. (2016). Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. *European Journal of Soil Science* 67: 276 - 284

Khorshidi, L. (2015). Soil-water interaction at high soil suction. A thesis submitted for Award of degree of Doctor of Philosophy of Colorado School of Mines, Colorado, USA. 148pp.

Kilasara, M. (2010). Selection and use of soil characteristics in digital soil mapping in Tanzania. In: *Proceedings of the 19th World Congress of Soil Science*. (Edited by Gilkes, R. and Prakongkep, N.)1 – 6 August 2010, Brisbane, Australia. 377 - 378pp.

Kovačević, M., Bajat, B., Gajić, B. (2010). Soil type classification and estimation of soil properties using support vector machines. *Geoderma* 154: 340 - 347

Lamorski, K., Pachepsky, Y., Sławiński, C. and Walczak, R. T. (2008). Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. *Soil Science Society of America Journal* 72(5): 1243 - 1247.

Lamorski, K., Slawinski, C., Moreno, F., Barna, G., Skierucha, W. and Arrue, J. L. (2014). Modelling Soil Water Retention Using Support Vector Machines with Genetic Algorithm Optimisation. *The Scientific World Journal*: 1-10

Li, H., Leng, W., Zhou, Y., Chen, F., Xiu, Z. and Yang, D. (2014). Evaluation Models for Soil Nutrient Based on Support Vector Machine and Artificial Neural Networks. *The Scientific World Journal* 2014, Article ID 478569: 1-7

Ließ, M., Schmidt, J., Glaser, B. (2016). Improving the Spatial Prediction of Soil Organic Carbon Stocks in a Complex Tropical Mountain Landscape by Methodological Specifications in Machine Learning Approaches. PLoS ONE: 11(4): e0153673.doi:10.1371/journal.pone.0153673

Mdemu, M. V. and Mulengera, M. K. (2002). Using pedotransfer functions (PTFs) to estimate soil water retention characteristics (SWRCS) in the tropics for sustainable soil water management: Tanzania case study. In: Proceedings of 12th International Soil Conservation Organisation conference on sustainable utilization of global soil and water resources. (Edited by Yuren, J. et al.), 26 - 31 May 2002, Beijing, China. 657 - 662pp.

Mdemu, M. V. (2015). Evaluation and Development of Pedotransfer Functions for Estimating Soil Water Holding Capacity in the Tropics: The Case of Sokoine University of Agriculture Farm in Morogoro, Tanzania. *Journal of Geography and Geology* 7(1): 1-9

Miháliková, M., Özyazıci, M. A. and Dengiz, O. (2016). Mapping Soil Water Retention on Agricultural Lands in Central and Eastern Parts of the Black Sea Region in Turkey. Journal of Irrigation and Drainage Engineering 142(12): 1-9.

Minasny, B., Hartemink, A. E., 2011. Predicting soil properties in the tropics. *Earth Science Reviews* 106: 52-62.

Mugabe, F. T., (2004). Pedotransfer functions for predicting three points on the moisture characteristic curve of a Zimbabwean soil. *Asian Journal of Plant Science* 3: 679 - 682.

Nemes, A., Rawls and W. J., Pachepsky, Y. A. (2006). Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. *Soil Science Society of America Journal* 70: 327 - 336.

Nguyen, P. M., Van Le, K., Cornelis, W. (2014). Using categorical soil structure information to improve soil water retention estimates of tropical delta soils. *Soil Research* 52: 443–452.

Nguyen, P. M., Pue, J. D., Van Le, K., Cornelis, W. (2015). Impact of regression methods on improved effects of soil structure on soil water retention estimates. *Journal of Hydrology* 525: 598–606.

Obalum, S.E and Obi, M.E. (2013). Moisture characteristics and their point pedotransfer functions for coarse-textured tropical soils differing in structural degradation status. *Hydrolological Processes* 27: 2721–2735.

Pachepsky, Y. A. and Rawls, W. J. (1999). Accuracy and reliability of pedotransfer functions as affected by grouping soils. *Soil Science Society of America Journal* 63: 1747 - 1757.

Raes, D., Steduto, P., Hsiao, T. C. and Fereres, E. (2009). AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: II. Main Algorithms and Software Description. *Agronomy Journal* 101(3): 438 – 447.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [http://www.R-project.org/] site visited 31/09/2016.

Schaap, M. G. (2005). *Models for indirect estimation of soil hydraulic properties. In: Encyclopedia of hydrological sciences*. (Edited by Anderson, M.). John Wiley & Sons, Ltd, New York. pp. 1145 – 1150.

Tóth, B., Weynants, M., Nemes, A., Makó, A., Bilas, G. and Tóth, G. (2015). New generation of hydraulic pedotransfer functions for Europe. *European Journal of Soil Science* 66: 226–238.

Tuller, M. and Or, D. (2001). Hydraulic conductivity of variably saturated porous media: Film and corner flow in angular pore space. *Water Resources Research* 37 (5): 1257-1276.

Twarakavi, N. C. K., Šimůnek, J., Schaap, M. G. (2009). Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters using Support Vector Machines. *Soil Science Society of America Journal* 73:1443 – 1452.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York, NY, USA. 319pp.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M. G. and van Genuchten, M.Th. (2010). Using Pedotransfer Functions to Estimate the van Genuchten– Mualem Soil Hydraulic Properties: A Review. *Vadose Zone Journal* 9: 795 – 820.

Vereecken, H., Schnepf, A., Hopmans, J. W., Javaux, M. Or, D., Roose, T., Vanderborght, J., Young, M. H., Amelung, W., Aitkenhead, M., Allison, S.D., Assouline, S., Baveye, P., Berli, M., Brüggemann, N., Finke, P., Flury, M., Gaiser, T., Govers, G., Ghezzehei, T., Hallett, P., Hendricks Franssen, H. J., Heppell, J., Horn, R., Huisman, J. A. , Jacques, D., Jonard, F., Kollet, S., Lafolie, F., Lamorski, K., Leitner, D., McBratney, A., Minasny, B., Montzka, C., Nowak, W., Pachepsky, Y., Padarian, J., Romano, N., Roth, K., Rothfuss, Y., Rowe, E. C., Schwen, A., Šimůnek, J., Tiktak, A., Van Dam, J., van der Zee, S. E. A. T. M., Vogel, H. J., Vrugt, J. A., Wöhling, T. and Young, I. M. (2016). Modeling Soil Processes: Review, Key Challenges, and New Perspectives. *Vadose Zone Journal*15(5): 1 - 57.

Wosten, J. H. M., Pachepsky, Y. A. and Rawls, W. J. (2001). Pedotransfer functions: bridging the gap between available soil data and missing soil hydraulic characteristics. *Journal of Hydrology* 251: 123 - 150.

Wosten, J.H.M., Verzandvoort, S.J.E., Leenaars, J.G.B., Hoogland, T., Wesseling, J.G. (2013). Soil hydraulic information for river basin studies in semi-arid regions. *Geoderma* 195–196:79–86.

Wu, C. H., Ho, J. M. and Lee, D. T.(2004). Travel Time Prediction with Support Vector Regression.*IEEE transactions on intelligent transportation systems* 5(4): 276 -281.

Young, M. D. B., Gowing, J. W., Hatibu, N., Mahoo, H. M. F., and Payton, R. W. (1999). Assessment and development of pedotransfer functions for Semi-Arid Sub-Saharan Africa. *Physics and Chemistry of the Earth – European Geophysical Society (B)* 24: 845–849.

CHAPTER FOUR

4.0 THREE-DIMENSIONAL MAPPING OF SOIL MOISTURE HOLDING CAPACITY WITH SOIL DEPTH FUNCTIONS AND MACHINE LEARNING

ABSTRACT

Soil moisture holding capacity (MHC) is highly variable and greatly influences agricultural productivity. Machine learning methods and soil depth functions (SDF) offer means for accurate and detailed characterization of lateral and vertical variability of MHC. This study examined the application of machine learning algorithms and soil depth functions for 3-dimensional mapping of MHC. Selected spatial ancillary data was subjected to principal component analysis as covariates for MHC prediction. Equal-area quadratic spline soil depth functions were fitted to model continuous vertical distribution of MHC data. Random forests (RF) and Cubist decision trees (CB) machine learning algorithms were trained on SDF fitted data to predict MHC with principal components of spatial covariates as predictors. Validation was performed at 3 measurement depths 15-cm, 45-cm, and 75-cm with mean error (ME) and root mean square error (RMSE) and R^2 as indices. Computations were performed in R-software. Ten principal components had eigenvalues > 1 with a cumulative variance > 70 %. Prediction accuracy was good with RMSEs ranging between 0.011-0.015 $cm^{-3}cm^{-3}$and R^2 between 36 - 81.4 %. Random forests had better accuracy than the Cubist decision trees. An RF-CB ensemble improves prediction accuracy. Observed results could be due to finer resolution of mapping covariates and learning ability of algorithms.

Keywords: MHC, Random forests, Cubist rules, Machine learning, Soil depth functions

4.1INTRODUCTION

Interest in detailed soil moisture information is high for modelling agricultural productivity in areas highly influenced by climate variability (Hengl *et al.*, 2017). Variability of soil moisture strongly influences the economic and environmental aspects of agricultural production systems (Junior *et al.*, 2014). Soil moisture exhibits high lateral and vertical variability (Chen *et al.*, 2015) with direct implications for water andnutrient management, erosion control and the sustainability of agricultural production systems (Junior *et al.*, 2014; Gray *et al.*, 2015).

Appropriate resource management actions require detailed characterization of lateral and vertical variation of soil moisture properties in form of three dimensional maps. However, this is limited by high soil sampling costs resulting in sparse data. Sparse data may lead to spurious representations of spatial structure of soil moisture characteristics and thus quantitative methods for preparation of detailed soil maps from this limited data are necessary. Digital soil mapping (DSM) represents a basis for quantification of soil properties through statistical and mathematical tools (Kempen, 2011; Mansuy *et al.*, 2014). DSM derives from empirical descriptions of the ‘SCORPAN’ (Soil attribute, Climate, Organisms, Relief, Parent Materials, Age, and spatial-locatioN) model (McBratney *et al.*, 2003), adapted from the ‘ClORPT’ (Climate, Organisms, Relief, Parent Materials and Time) soil state factor model (Jenny, 1941). It benefits from availability of spatial environmental data like remote sensing data and digital elevation models, as soil formation ancillary covariates to develop predictive relationships from limited point observations (Malone *et al.*, 2009).

Various statistical and mathematical methods have been explored to develop quantitative soil-covariate relationships for prediction or interpolation of soil properties. Geostatistical modelling (e.g. ordinary kriging, co-kriging, and regression-kriging) has been most widely applied (Junior *et al*., 2014). Geostatistical modelling assumes normal distribution of data (Heuvelink, 2014), and leads to less reliability of results with non-normally distributed data (Kavianpoor *et al*., 2012), in absence of spatial autocorrelation (Lark, 2012) or limited data points (Heuvelink, 2014). Machine learning methods (e.g. random forests, cubist decision trees, support vector machines) have also been applied in mapping soil properties (e. g. Kovačević *et al.*, 2010; Mansuy *et al*., 2014; Ließ *et al*., 2016; Hengl *et al*., 2017). No assumptions on distribution class of data is made with machine learning algorithms (MLAs) (Ustner *et al*., 2015), which transcends the drawbacks of geostatistical modelling for predictive mapping of soil MHC. Kovačević *et al*. (2010) and Hengl *et al*. (2017) found MLAs to be suitable for soil predictive mapping. The method applied for predictive mapping depends on a variety of factors like available soil data and environmental covariates, size and environmental characteristics of the area mapped, the processing time, ease of model implementation and result interpretation, as well as the desired mapping accuracy (Junior *et al*., 2014).

Approaches for 3-D mapping of soil properties have mostly employed hybrid methods integrating geostatistics and soil depth functions to map variability of soil properties in 3-D (Malone *et al.*, 2009; Kempen, 2011; Veronesi *et al*., 2012; Adhikari *et al*., 2013). Soil depth functions are models fitted to the discrete measurements of soil attribute to represent a continuous vertical distribution of soil attributes in a profile. Polynomial splines (Malone *et al.*, 2009; Veronesi *et al.*, 2012; Adhikari *et al.*, 2013; Mulder *et al.*, 2016), exponential decay functions (Kempen, 2011), power and logarithmic functions (Liu *et al.*, 2013) have been used as soil depth functions. Quadratic spline functions have been documented as the most accurate method for fitting soil data in a vertical continuum of the soil profile (Malone *et al.*, 2009; Veronesi *et al.*, 2012; Liu *et al.*, 2013). This approach to the 3-dimensional problem does not explicitly consider the 3-D spatial dependence (Liu *et al.,* 2013) and insufficiently represents the support over which the soil data is collected (Orton *et al.*, 2016). Therefore, the objective of this research was to evaluate random forests and cubist decision trees for 3-dimensional mapping of soil moisture holding capacity.

4.2MATERIALS AND METHODS

4.2.1 Study area

This study was undertaken in Ilakala village in Kilosa District – Tanzania. It has a total area of about 44 km^2 and altitude between 546 - 1000 m. It borders Mikumi National Park to the South. Western and Northern peripheries are spanned with hilly relief in proximity to the Rubeho Ranges, predominantly overlain with unprotected Miombo woodlands vegetation. Arable agriculture predominates in the Southern areas. The area experiences a dry sub-humid climate with a mean annual rainfall of 500 – 800 mm. It is located within latitudes 7^{o}5’ 30” S and 7^o 9’ 30” Sand longitudes 36^o 50’ 30” E and 36^o 57’ 30” E (Fig. 4.1).

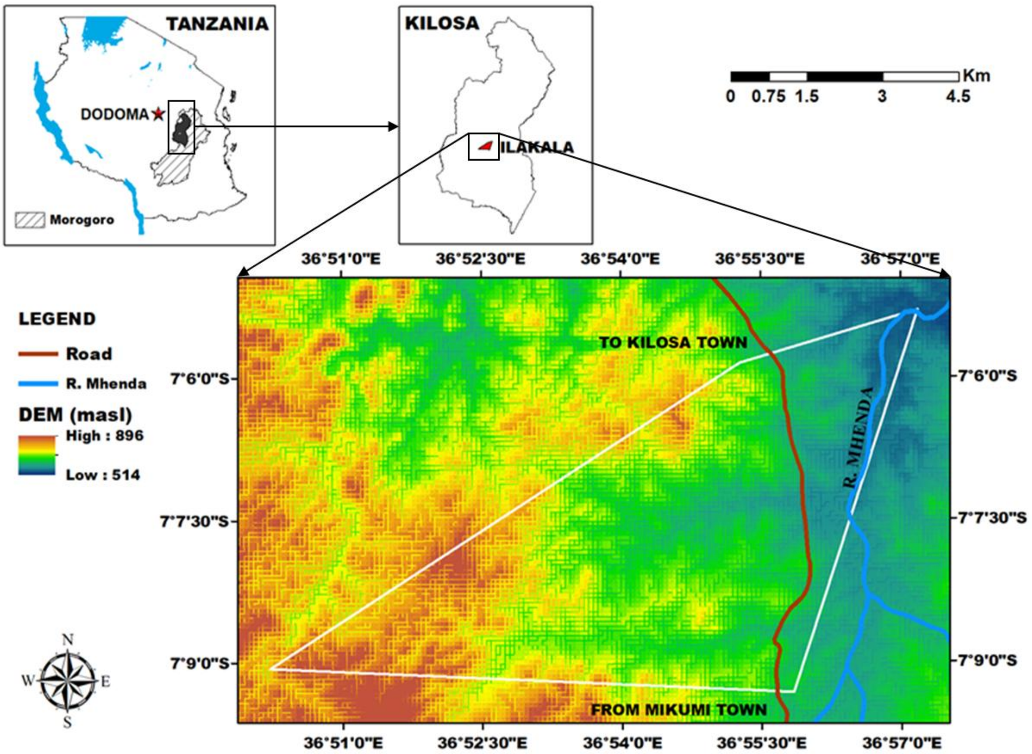

Figure 4.1: Location of study area

4.2.2 Soil dataset

Soil data was collected from 100 locations through a stratified random sampling design. Samples were collected at depth intervals of 0-30cm (15-cm), 30-60 cm (45-cm), and 60-100 cm (75-cm). Numbers in the brackets indicate the specific sampling depths. Four locations at the 60-100 cm depth interval weren’t sampled due to rockiness. A bulk soil sample of about 500 g was excavated and duplicate soil cores samples of 100 cc taken at each depth interval. Bulk soil samples were air-dried and sieved for subsequent analysis for soil particle size distribution and organic carbon. Particle size fractions were measured by the Bouycous hydrometer method (Gee and Bauder, 1986). Organic carbon was measured by the Walkley and Black wet oxidation method (Nelson and Sommers, 1982). Moisture retention and bulk density were measured from the soil core samples. Moisture retention measurements at field capacity (FC) and wilting point (WP) were made using a pressure plate apparatus at suctions of 30 kPa and 1500 kPa, respectively. Soil moisture holding capacity (MHC) was taken as the difference of moisture at FC and WP. After measurements of wilting point, the soil core samples were oven-dried at 105 ^{o} C for 24 hours for estimation of bulk density (Blake and Hartge, 1986). To have a complete data-set for all the measurement depths, the missing data from the 4 sampling locations was imputed as a mean of the other measured points.

4.2.3 Descriptive Statistics

Box-plots for the measured soil physicochemical properties were constructed to assess the variation of soil properties across measured soil depths intervals. Exploratory statistics were used to determine the mean, minimum (MIN), maximum (MAX), standard deviation (SD), coefficient of variation (CV) and Shapiro Wilkinson test of the measured MHC data. The Shapiro-Wilkinson test is a measure of distribution of data for normality.

4.2.4 Auxiliary environmental variables

The modelling framework is hinged on auxiliary information from spatial covariates.Table 4.1 shows the selected spatial covariates used in this study.Primary spatial covariate layers selected for this study were soil type (SOILS), geology (GEO), and digital elevation model (DEM).

The soils layerwas developed within the framework of Trans-Sec Project (Trans-Sec, 2017). It is composed of eight categorical soil mapping units corresponding to five major indigenous soil types (*Kichanga*, *Mfinyanzi*, *Ngunja*, *Tifu-Tifu*, *Wakitope-Mweusi*), and three soil associations composed of combinations of indigenous soil types (*Mfinyanzi+Kichanga, Ngunja+Kichanga, Tifu-Tifu+Kichanga*) (Kaingo and Tumbo, 2016). Appendices 1 to 5 show the field data for the indigenous soil units and indicates therespectiveinternational soil referencename(WRB, 2015).

Geology data was from Geological Survey of Tanzania and it consisted of four categorical geologic units (XAB, RAL, MBG and NTS). The lithology of these geologic units is described in Appendix 6.The DEM was of the 30-m Shuttle Radar Topography Mission (SRTM) provided by United States Geological Survey (USGS) (USGS, 2017).

Other spatial covariates were derived from the DEM using the SAGA processing toolbox in QGIS 2.16 software (QGIS Development Team, 2016). DEM-derivatives included Slope (SLP), Aspect(ASP), SAGA topographic wetness index (STWI), Topographic Wetness Index (TWI), Topographic Position Index (TPI), Length-slope factor (LSF), Altitude Above Channel Network (AACN), Planar Curvature (PLC), Profile Curvature (PRC), Multi-Resolution Valley Bottom Flatness Index (MRBF), and Multi-Resolution Ridge-Top Flatness Index (MRTF). The importance of these DEM-derivatives to soil-landscape modelling has been highlighted by different authors (e. g. Moore *et al*., 1993;Böhner and Selige, 2006;Buchanan *et al*., 2014; Viloria*et al*., 2015;Orton*et al*., 2016).

Table 4.1: Spatial Environmental Covariates

SHORT	Descriptive name	RESOLUTION	SOURCE
AACN	Altitude above channel network	30-m	DEM-Derived
ASP	Aspect	30-m	DEM-Derived
DEM	Digital Elevation Model	30-m	USGS (2017)
GEO	Geology	1:125 000	Geological Survey of Tanzania (2016)
LSF	Slope-length factor	30-m	DEM-Derived
MRBF	Multi-resolution valley-bottom flatness index	30-m	DEM-Derived
MRTF	Multi-resolution ridge-top flatness index	30-m	DEM-Derived
PLC	Plan curvature	30-m	DEM-Derived
PRC	Profile curvature	30-m	DEM-Derived
SOILS	Soil Mapping units of Ilakala Village	1:12 500	Kaingo and Tumbo (2016)
STWI	SAGA Topographic Wetness Index	30-m	DEM-Derived
TPI	Topographic Position Index	30-m	DEM-Derived
TWI	Topographic Wetness Index	30-m	DEM-Derived

4.2.5 Principal component analysis (PCA)of auxiliary spatial covariates

Principal component analysis (PCA) was performed on a matrix of spatial covariates to extract the most important information from the data, and for dimensional reduction for analysis of the structure of the observation and variables. Principal component analysis is a multivariate statistical technique used to generate latent variables known as principal components (PCs) for analysing variance structure of data as weighted linear combinations of original variables (Young and Pearce, 2013; SAS, 2017). All spatial covariate layers developed in QGIS software were resampled or calculated at a spatial resolution of 30-m. Spatial covariate data was then imported into the R-software environment (R Core Team, 2016) and the ‘spatial predictive components’ function (spc) of the R-GSIF package (Hengl, 2016) was used to calculate the PCs of the covariate data. Miscellaneous R-base functions were used to calculate the eigenvalues and variance contribution of the PCs from the GSIF output. Principal components with eigenvalues greater than 1 were retained for further analysis following the Kaiser criterion (Young and Pearce, 2013).

4.2.6 3-Dimensional mapping of MHC

Two mapping approaches were considered i.e. a scenario with a complete set of measured MHC data (DSM-A) and a scenario combining estimated and measured MHC data (DSM-B). The principle of saving sampling costs guided the need to evaluate the prediction efficiency of combining estimated and measured data in predictive mapping. Support vector machines pedo-transfer functions (SVM-PTF) earlier developed for the study area (Chapter Three of this dissertation), were applied for estimation of MHC with 50 % of the sampling points. Sampling points for PTF-estimation were selected using k-means clustering with the ‘CLUSTER’ directive of Genstat-15 software (VSNi, 2017). Clustering was implemented in such a way that an estimated and measured point occurred in adjacency within the same geostrata.

The work-flow described in this paragraph consistently applied to both mapping scenario DSM-A and DSM-B. Equal-area quadratic spline soil depth functions were used to fit the measured MHC data at 1-cm vertical resolution across the 1-m depth interval with knots at 0, 15, 45 and 75-cm. Data of the most significant principal components was extracted for each sampling point and combined with the spline-fitted MHC data to generate a regression matrix. Machine learning algorithms (Random forests (RF) and cubist rules (CB)) were trained on 60 % of randomly selected data points from the regression matrix (n = 6000) consisting of key PCs and soil depth as predictors. The *randomForest* (Liaw and Wiener, 2002) and *Cubist* (Kuhn *et al.*, 2016a) packages of R-software were used for training RF and CB algorithms, respectively, through the *caret* R-package (Kuhn *et al.*, 2016b). The model construct fed into the RF and CB algorithms had MHC as the dependent variable with PCs and soil depth as predictors. Predictions were subsequently performed using the trained models. Predictive mapping using an RF-CB ensemble as a weighted average of coefficients of determination was also evaluated. Validation was performed with the remaining 40 % of data points using the RMSE, R^2, and ME as prediction accuracy indices.

$$\mathbf{ME} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}[\mathbf{y} - \hat{\mathbf{y}}\,]$$

$$\mathbf{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}[\,\mathbf{y} - \hat{\mathbf{y}}\,]^{\mathbf{2}}}$$

$$\mathbf{R^2} = \left(\mathbf{1} - \frac{\sum_{i=1}^{n}[y - \hat{y}]^2}{\sum_{i=1}^{n}[y - \bar{y}]^2}\right)^{\mathbf{2}}$$

Where y is the measured moisture content, ŷ is the predicted moisture content and ȳ is the mean of the measured moisture content.

4.3RESULTS AND DISCUSSION

4.3.1 Descriptive statistics

Figure 4.2 shows the variation of the measured soil properties with soil depth. Organic carbon (OC) content was highest in the top soil and gradually reduced with soil depth. Variability and value of bulk density was highest at the 0-30 cm depth interval. Median BD values of for the 30-60 cm and 60-100 cm depth were similar. Proportion of soil particle sizes across all depths followed the order sand> clay> silt. Sand content exhibited a decrease with increased soil depth. Moisture content was lowest for the 0-30 cm depth interval for both the field capacity and wilting point. Least variability of moisture content was observed at the furthest depth (60-100 cm).

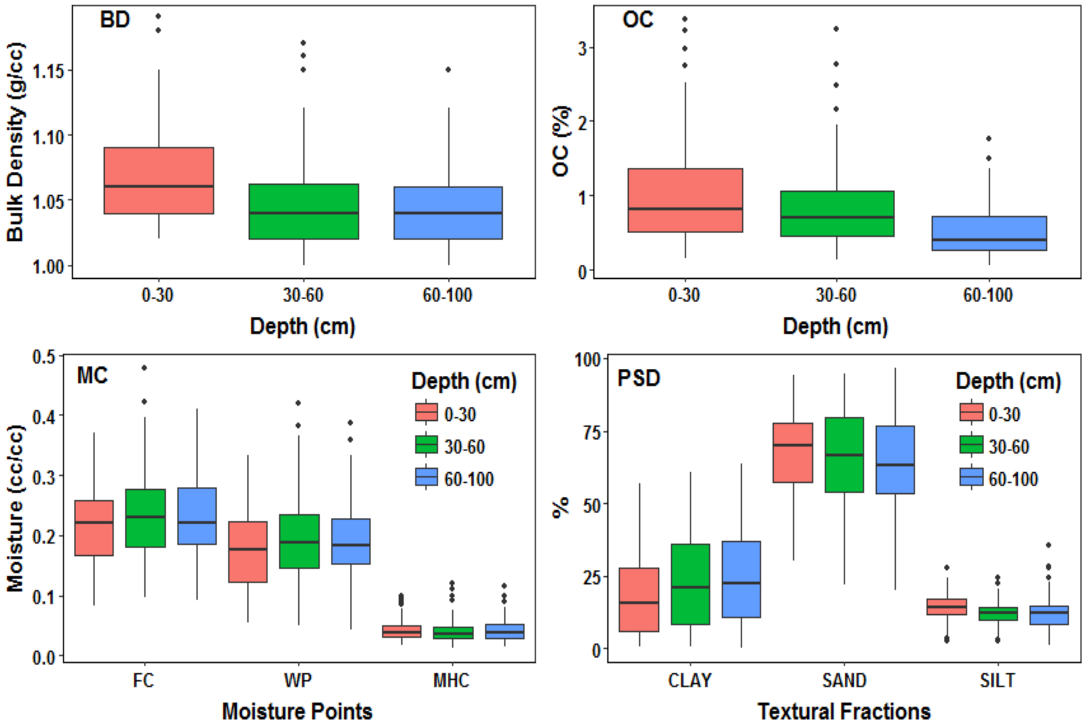

Figure 4.2: Boxplots showing variation of soil properties with depth

Table 4.2 shows the descriptive statistics of MHC at the various depths. Moisture holding capacity was highest at the 60-100 cm depth with mean value of 0.043 $cm^{-3}cm^{-3}$. However, mean MHC at the other depths is not significantly different. The MHC is slightly more uniform at the top depth (CV= 40 %) compared to the lower depths. The highest variability of MHC was at the 30-60 cm depth. Shapiro-Wilkinson normality test scores (S-W) were significant indicating that MHC data at all depths do not fit a normal distribution.

Table 4.2:Descriptive statistics of moisture holding capacity across depths

DEPTH (cm)	MIN	MEAN	MAX	SD	CV (%)	S-W
0-30	0.016	0.042	0.1	0.017	40.028	0.88***
30-60	0.012	0.04	0.121	0.019	48.214	0.867***
60-100	0.014	0.043	0.115	0.02	46.383	0.909***

4.3.2 PCA of spatial covariates

A total of 24 PCs were derived from the original 14 spatial covariates used in the study. Figure 4.3 depicts the cumulative variance of the derived PCs from the spatial covariates. Principal component 1 (PC1), PC2, and PC3 individually account for about 19 %, 11 %, and 9 % of the variance in the spatial covariates, respectively. Cumulative variance of the last three PCs (PC22, PC23, and PC24) was marginal (<< 1 %) and perhaps represents noise within the spatial covariate data.

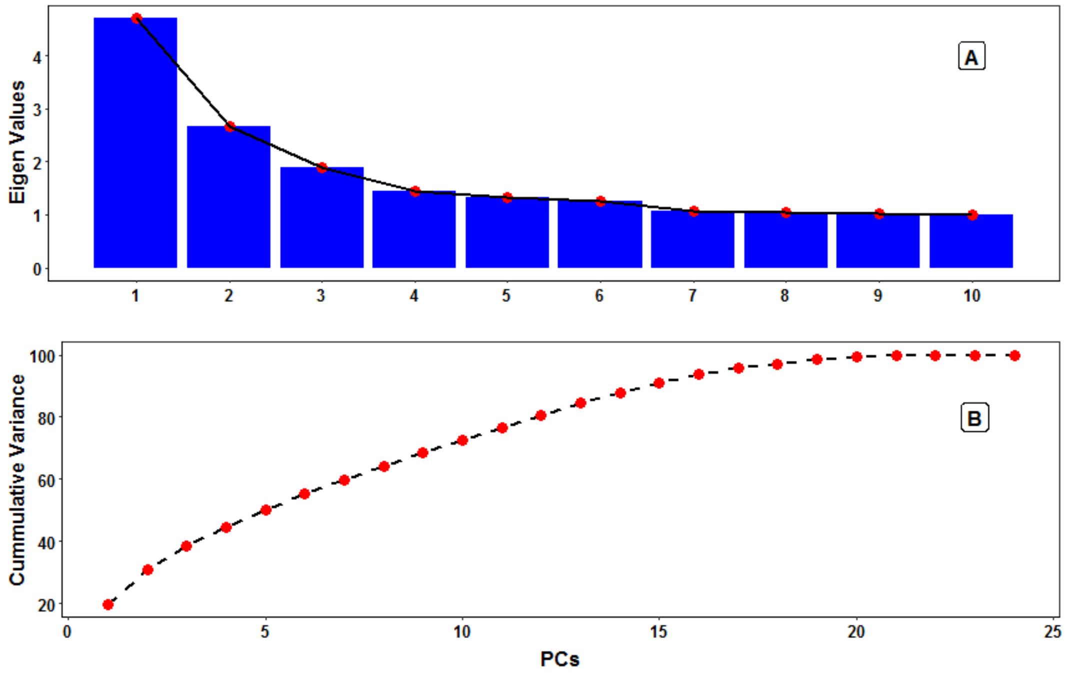

Figure 4.3: Eigenvalues and Cumulative Variance of Spatial Covariates Principal Components (PCs)

Principal components 1 to 10 (PC1-PC10) had eigenvalues greater than 1 as shown in Fig. 4.3A. These satisfy Kaiser’s criterion (eigenvalue > 1), implying that PC1 to PC10 were significant for subsequent analyses (Young and Pearce, 2013; Bingol *et al.*, 2013). Iezzoni and Pritts (1991) highlighted that PCs with an eigenvalue > 1, are inherently more informative than any single original covariate taken alone. The ten extracted PCs (PC1-PC10) accounted for over 70 % of the total variance in the spatial covariate data (Fig. 4.3B). Principal components with a cumulative variance of about 80 – 90% suitably substitute original variables (Bingol *et al.*, 2013), although a cumulative variance of 70 % is also considered permissible for model constructs (SAS, 2017). A variance of 80 % would have been achievable with the inclusion of PC11 and PC12 but their eigenvalues fell below the Kaiser threshold. Iezzoni and Pritts (1991) argue that PCs with eigenvalue < 1 merit inclusion where physical meaning can be attributable to the PCs. However, for model parsimony, inclusion of PC 11 and PC 12 was decided against. Comparable to this study, Wang et al. (2012) found that inferential tests yielded PCs that explained cumulative variance of between 69 – 75 %.

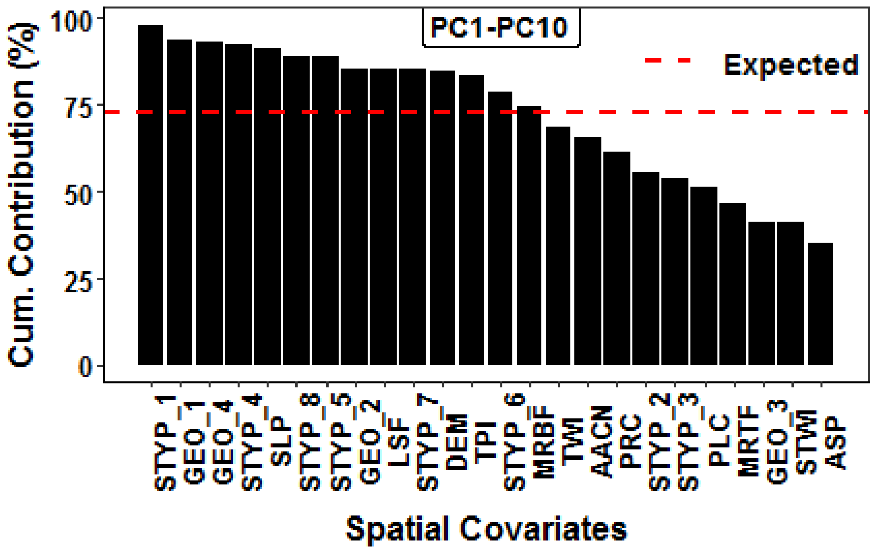

Figure 4.4: Pooled variance of spatial covariates across PC1-PC10

Figure 4.4 shows the pooled variance of the individual spatial covariates across the retained PCs (PC1-PC10). Soil unit 1 (STYP_1) had the highest pooled variance with a cumulative variance of about 90 %. Aspect (ASP) had the least pooled variance on the PCs with a cumulative variance of less than 40 %.Six of the spatial covariates with the highest cumulative variance across PC1-PC10 are illustrated in Fig. 16. These were soil types (SOILS), geology (GEO), slope (SLP), slope length factor (LSF), elevation (DEM) and topographic position index (TPI). Legend item 1 of the ‘SOILS’ layer (Fig. 4.5) represents ‘STYP_1’ (Fig. 4.4) and corresponds to indigenous soil type *Kichanga*. It is a Hyperdystric Cambisol(loamic, ochric)(Appendix 1) with sand proportions ≥ 70 % (Kaingo and Tumbo, 2016). It had the largest areal coverage which possiblycaused the observed cumulative variance. The other legend items for the 'SOILS' layer are:*Mfinyanzi* (2), *Mfinyanzi+Kichanga* (3), *Ngunja* (4), *Ngunja+Kichanga* (5), *Tifu-Tifu* (6), *Tifu-Tifu+Kichanga* (7), and *Wakitope-Mweusi* (8). Correspondingly, the largest geologic unit (legend item 1) for the geology layer (GEO) (Fig. 4.5), had the 2nd highest cumulative variance (Fig. 4.4, GEO_1). This corresponds to the geologic unit code XAB (Appendix 6). The XAB lithologic unit is primarily composed of meta-sedimentary rocks with soils formed frombiotite gneissparent material. The geologic legend items are: RAL (2), MBG (3), and NTS (4).

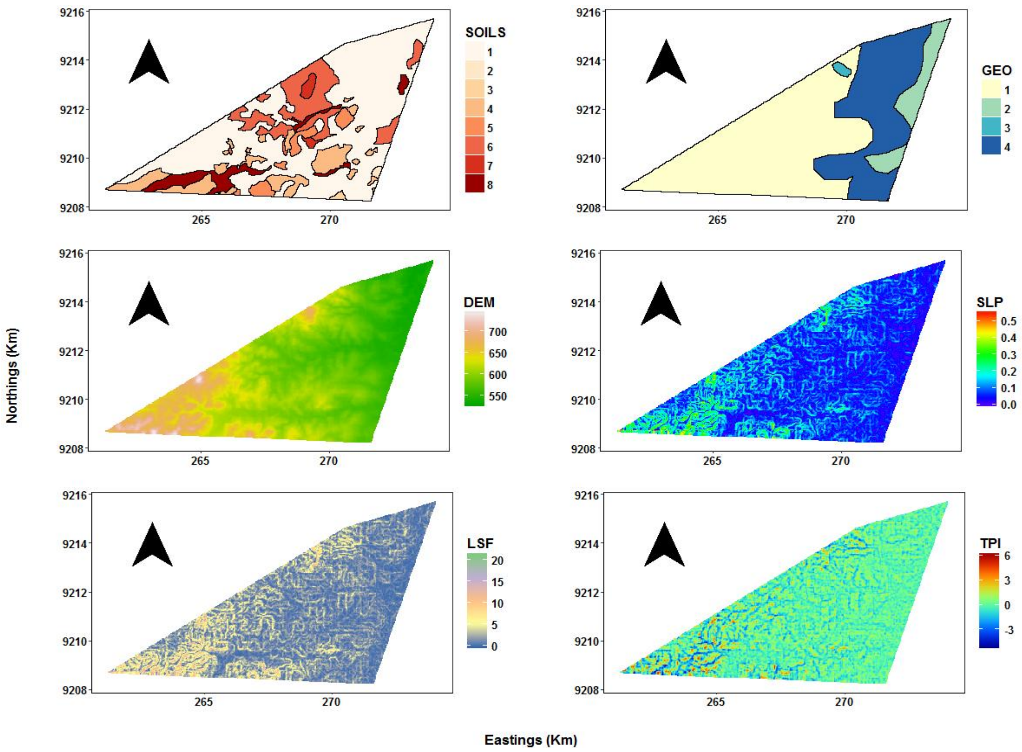

Figure 4.5: Spatial covariates with highest cumulative variance contribution to PC1-PC10

An overall outlook is that soil type and geology had the most influence on the PCs (Fig. 4.4). Other studies (Adhikari *et al*., 2014; Gray *et al*., 2016; Hengl *et al*., 2017) have emphasized the relative importance of the soils and geology covariates in spatial prediction of soil properties. Hengl *et al.* (2017) observed distinctive patterns of lithological classes in output prediction maps for texture and coarse fragments. For soil hydrological properties like MHC, particularly, lithological structure or parent materials represented by the geology covariate wield strong influence (Kodešová *et al*., 2009; Brocca *et al*., 2012; Gray *et al*., 2016). Mineral composition and degree of alteration of parent materials determines grain size of soils (Kodešová *et al*., 2009) influencing soil pore architecture with control on the soil moisture retention properties. Parent material high in silica content, often contains high quartz resulting in coarse sandy soils. Composition of underlying lithological structure for both the soil type and geology covariates was quartz (Kaingo and Tumbo, 2016) resulting in the observed sandy soils spanning a large extent of the study area (Fig. 4.5).

4.3.3 3-D mapping with DSM-A scenario

Figure 4.6 shows the prediction maps of MHC for the DSM-A mapping scenario at 15-cm, 45-cm and 75-cm depth. Highest predictions of MHC were concentrated in the eastern and south-western areas of the study area. High predictions in the south-western area correspond to a region with relatively high vegetative cover. High vegetative cover leads to accumulation of soil organic carbon which enhances the moisture retention properties of the soil (Kodešová *et al*., 2009; Ließ *et al*., 2016). Map predictions for MHC were generally higher for the 45-cm layer. The lowest

MHC map predictions are apparent for the 15-cm depth. Textural distribution in the horizons is assumed to be responsible for this observed pattern. Upper horizons generally had a coarser textural fraction, with finer textural fractions in the mid-intervals and a slightly fine to coarse fraction at deeper horizons towards the C-horizons. Spatial prediction patterns of the maps are similar for all algorithms (RF, CB and AVG).

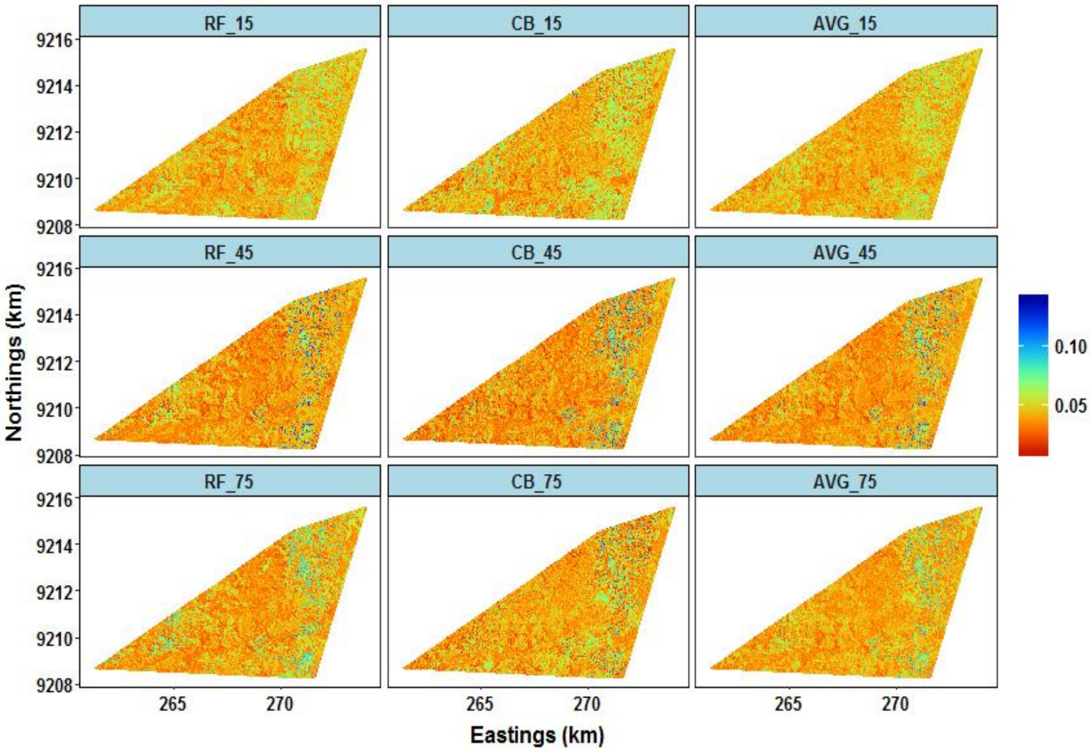

Figure 4.6: Maps of MHC with Random Forest (RF), Cubist Rules (CB) and Ensemble approach (AVG) at 15 cm, 45 cm and 75 cm depth

Figure 4.7 shows the distribution of the training and validation points used for development and assessment of mapping framework at 15 cm, 45 cm and 75 cm depth. These points were objectively selected through a probabilistic sampling design and thus are suitable for statistical validation of maps without bias. Training points ranged from 58-60 while validation points ranged from 37-40. It is important to note that the validation data from these locations is independent and was not used in training the RF and CB algorithms. Brus *et al.* (2011) pointed out that unbiased and valid prediction map quality estimates are best achieved using independent data not used in modelling, and selected by probability sampling.

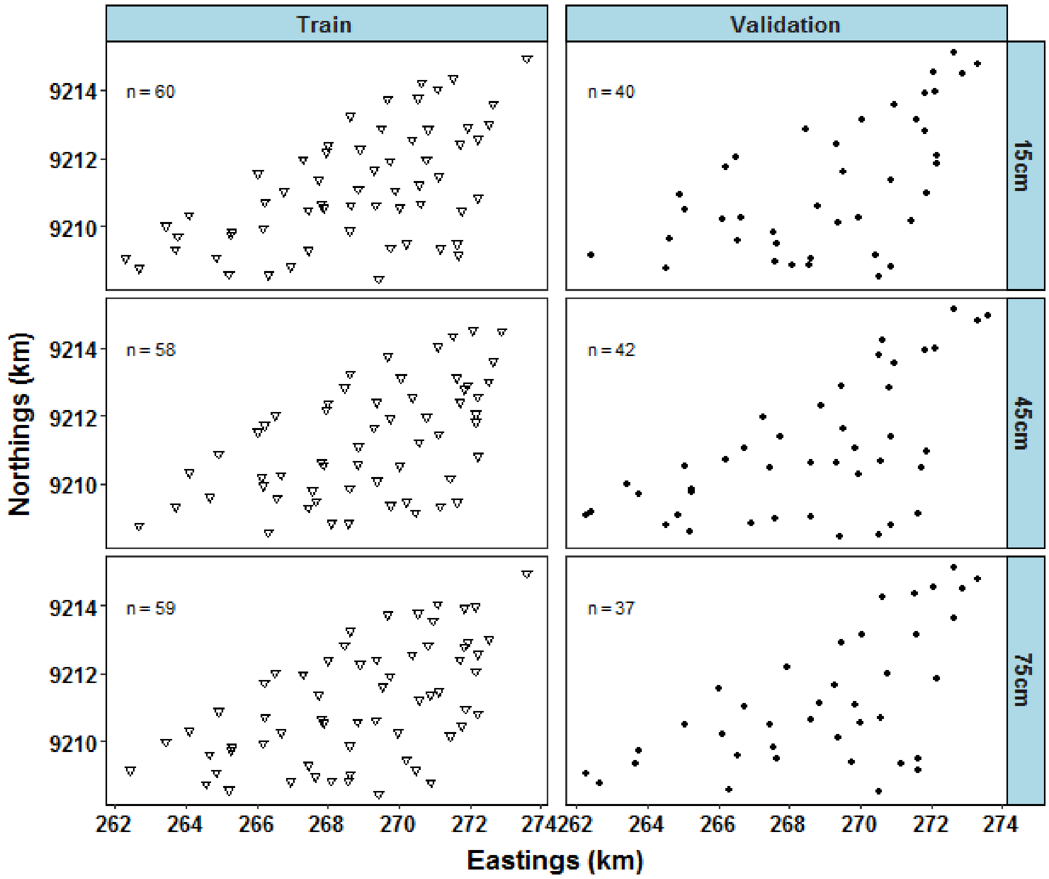

Figure 4.7: Location of training and validation points at 15-cm, 45-cm and 75-cm depths

Validation results of the prediction maps are given in Table 4.3. Overall observation is that MEs are very close to zero; implying limited bias in spatial prediction of MHC. The MLAs though tend to overestimate MHC at the 15-cm depth (MEs <0).

The cubist algorithm (CB) overestimates MHC at 75-cm depth with ME < 0. Random forest algorithm (RF) appears to have the least MEs across the depths. Mean error of the CB-RF ensemble (AVG) tends to be lower than the ME for CB. Zero MEs of RF and AVG at 45-cm and 75-cm depth, respectively, could be due to cancelling out of MHC values straddling above or below the line of best fit (Fig. 4.8). Observed RMSEs for RF, CB and AVG, ranged from 0.011-0.015 cm^3cm^{-3}, 0.011-0.013 cm^3cm^{-3}, and 0.008-0.013 cm^3cm^{-3}, respectively. Improvements in RMSE by applying MLAs for map prediction have been reported and attributed to better statistical learning abilities of the MLAs for non-linear soil data. RMSE for the 45-cm depth were lowest compared to other validation depths maps. The RMSE values of the CB-RF ensemble (AVG) were the lowest across all the validation depths. This suggests higher mapping accuracy of CB-RF ensemble prediction than mapping accuracy with RF or CB models alone, likely due to reduction of overshooting effects from either MLA (Hengl *et al*., 2017).

Table 4.3: Validation statistics forMHC mapping with DSM-A scenario

MLA	ME (cm^3cm^{-3})			RMSE (cm^3cm^{-3})			R^2 (%)		
	15cm	45cm	75cm	15cm	45cm	75cm	15cm	45cm	75cm
RF	-0.001	0.000	0.001	0.011	0.010	0.015	41	74.3	39.667
CB	-0.003	0.002	-0.001	0.012	0.011	0.013	36	65.1	55.747
AVG	-0.002	0.001	0.000	0.010	0.008	0.013	45.4	81.4	53.637

The R^2 results were between 36 - 81.4 %. Results indicate that R^2 values were highest for the 45-cm depth and overall lowest for the 15-cm depth except for the RF algorithm where the 75-cm depth returned the lowest R^2. Validation results as reflected by RMSE and R^2 broadly follow variability in MHC data as shown by the coefficient of variation (Table 4.3).

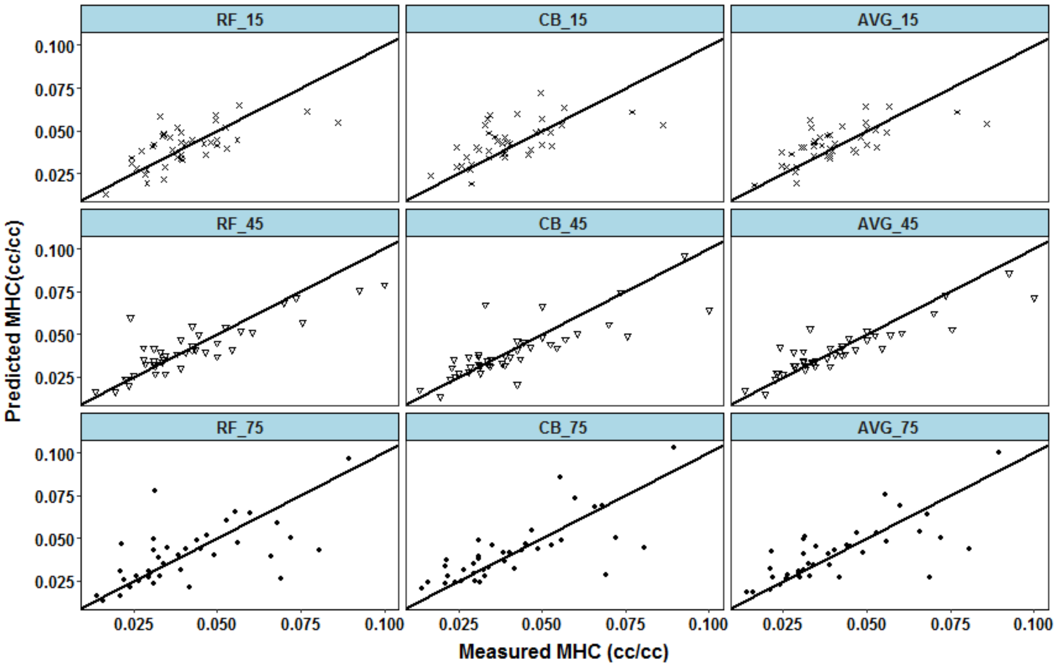

Figure 4.8: Unit plots for validation of MHC prediction maps

4.3.4 3-D mapping with DSM-B scenario

Even proportions of measured and PTF-predicted soil MHC data were combined for modelling 3-D prediction maps. Figure 4.9 shows the spatial distribution of locations with observed and PTF-estimated data. Locations were clustered in such a way that a measurement and PTF-estimated point occurred in each geostrata with adjacency. Prediction maps of the MLAs under the DSM-B scenario are shown in Fig. 4.10. Spatial prediction patterns for DSM-B were similar to the DSM-A scenario (Fig. 4.6) - though the signal seems lower especially for the south-western region (Fig. 21).

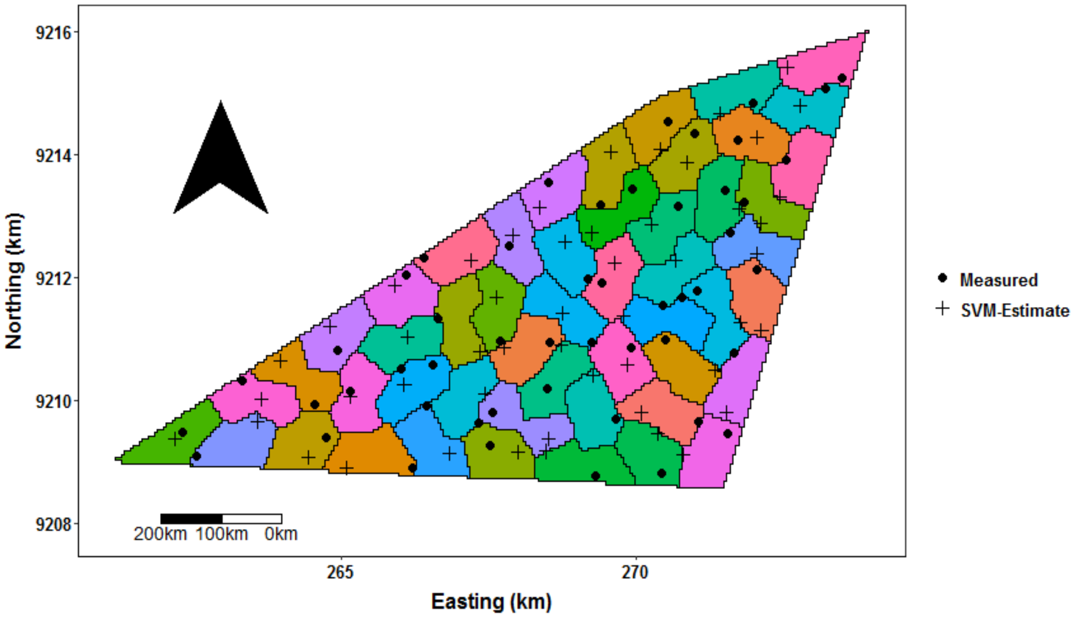

Figure 4.9: Distribution of points with SVM-PTF estimated MHC

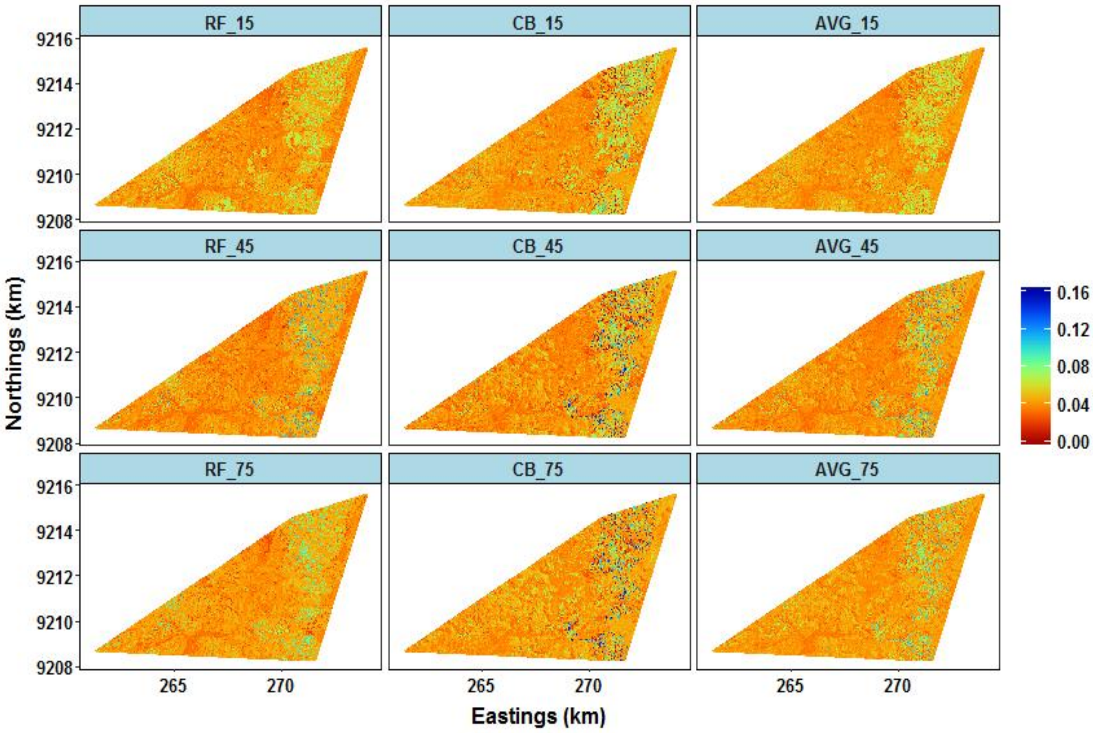

Figure 4.10: Maps of MHC with Random Forest (RF), Cubist Rules (CB) and Ensemble approach (AVG) at 15 cm, 45 cm and 75 cm depth

Generally, the cubist algorithm (CB) overestimates MHC at all depths (15-cm, 45-cm, and 75-cm) with MEs < 0 (Table 4.4). Random forest algorithm (RF) tends to underestimate for the 15-cm depth (ME > 0) while the CB-RF ensemble (AVG) overestimates at the 75-cm depth (ME < 0) (Table 4.4). The R^2 values ranged between 19-55% (Table 4.4) and were generally lower than the DSM-A scenario with the complete measured dataset (Table 4.3). The highest R^2 values were registered with the prediction maps for the 45-cm layer. These results are better than those reported by Malone *et al.* (2009) (8 - 29%). The difference between these results could be explained by the approaches applied in the study. Malone et al. (2009) used only PTF-estimates for mapping MHC while both measured and PTF-estimates were combined in this study.

Table 4.4: Validation statistics for MHC mapping with DSM-B scenario

MLA	ME (cm^3cm^{-3})			RMSE (cm^3cm^{-3})			R^2 (%)		
	15-cm	45-cm	75-cm	15-cm	45-cm	75-cm	15-cm	45-cm	75-cm
RF	0.001	0.000	0.000	0.012	0.013	0.015	30.7	55.0	33.8
CB	-0.001	-0.001	-0.001	0.014	0.014	0.015	19.0	48.2	34.2
AVG	0.000	0.000	-0.001	0.011	0.013	0.015	32.1	54.3	35.8

The RMSEs ranged from 0.011 cm^3cm^{-3} to 0.015 cm^3cm^{-3} (Table 4.4). The RMSEs were marginally higher than in the scenario with completely measured data (Table 4.3), implying that mapping accuracy decreased with inclusion of PTF-estimated data. However, the results are better than for the study by Miháliková *et al.* (2016) who reported RMSEs in the range of 0.035 – 0.092 cm^3cm^{-3}. This disagreement in results could be due to a multifold of factors viz differences in sizes of datasets, model used and scale of the mapping areas. Relative decrease in prediction accuracy was between 9 – 63 % (Table 4.5), with the 45-cm depth prediction maps showing the highest decrease in accuracy. This is most likely linked to error propagation from the PTF-estimated MHC data. Malone *et al.* (2009) reported a propagation of error into the spatial model due to use of extrapolated spline estimates. Machine learning algorithms have also been reported to be highly sensitive to errors in data (Hengl *et al.*, 2017). Therefore, using estimates of MHC enhanced the uncertainty within the MLAs during the model training phase resulting in the larger order of difference between predicted MHC and observed MHC values.

Table 4.5: Relative RMSE of the MLAs with inclusion of PTF-estimated data

DEPTH	AVG (%)	CB (%)	RF (%)
15-cm	-10.00	-16.67	-9.09
45-cm	-62.50	-27.27	-30.00
75-cm	-15.38	-15.38	0.00

There is indication that the spatial prediction patterns of MHC depend on the spatial patterns of input covariates. Misclassification error associated with spatial covariates especially categorical covariates (e.g. Soil type and geology) could have also resulted in a mismatch of spatial relationships with the dependent variable (Stoorvogel *et al.*, 2009; Hengl *et al.*, 2014).

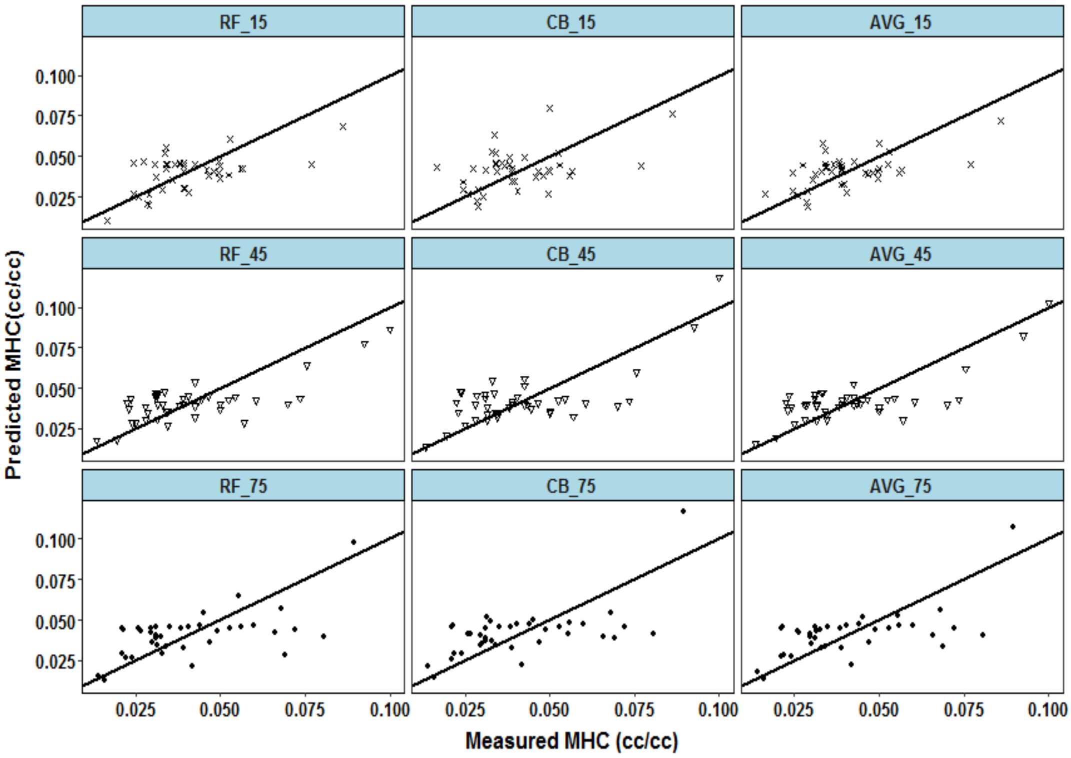

Figure 4.11: Unit plots for maps generated with PTF-predicted MHC

Error propagation limits the utility of prediction maps and might influence wrong decisions (Heuvelink, 2014). Although Hengl *et al.* (2014) vouch for application of more sophisticated statistical modelling as a mitigating approach, enhanced complexity of quantitative models might only provide limited improvements in prediction accuracy (Mansuy *et al.*, 2014). The MLAs are data-driven and as such only a simulation of observed data. The envisaged improvements in prediction accuracy with more model complexity will inevitably plateau. More reliable map predictions could perhaps be achieved with higher sampling density and better configuration of data points (Minasny *et al.*, 2007; Lark, 2012; Junior *et al.*, 2014). This though might be prohibitive due to costs and a trade-off has to be reached about the ‘optimal’ error to accommodate.

Spatial accuracy of the covariates needs to be improved more so for qualitative/discrete spatial covariates such the soil type and geology layers. Poor accuracy of categorical spatial covariates arises due to high short-scale spatial variation of the properties combined with a low sampling resolution. Improving thematic accuracy of the categorical classes needs more precise delineations especially where the target mapping resolution is smaller than the scale of the input spatial covariate. High resolution remote sensing data products like the TanDEM-X WorldDEM^{TM} offering a resolution of less than 12m (DLR, 2016a; 2016b), increasing affordability of technologies such as proximal sensing (Viscarra-Rossel *et al.*, 2011; Ji *et al.*, 2015) and low-altitude unmanned aerial vehicles (UAVs) (Crommelinck *et al.*, 2016) offers a viable pathway for high-order thematic classification for improved prediction map accuracy. The UAVs have potential to generate detailed DEM data with pixels of ground resolution of as high as 5 cm which could comprehensively improve representation of short scale variations of soil properties.

4.3CONCLUSIONS

A framework has been evaluated for 3-D mapping of soil moisture holding capacity using machine learning and soil depth functions. Random forests are better for predictive mapping than Cubist rules. An ensemble of CB and RF further improves accuracy of predictions. Predictive mapping using a combination of measured and PTF-predicted MHC data decreases accuracy. Using principal components as predictors returns good estimates of MHC. Though the predictions are 2-D surfaces, predictions of MHC layers can reliably be made at any depth of interest across the continuum of depth interval within which the algorithms were trained. Rendering could then be achieved within software libraries that accommodate 3-D plots like the plotKML R-package.

4.3REFERENCES

Adhikari, K., Bou Kheir, R., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A. B. and Greve, M. H. (2013). High-Resolution 3-D Mapping of Soil Texture in Denmark. *Soil Science Society of America Journal* 77(3): 860 – 876.

Adhikari, K., Minasny, B., Greve, M. B., and Greve, M. H. (2014). Constructing a soil class map of Denmark based on the FAO legend using digital techniques. *Geoderma* 214–215: 101–113.

Bingöl, D., Ümit Ay, U., Bozbas, S. K., Uzgören, N. (2013). Chemometric evaluation of the heavy metals distribution in waters from the Dilovası region in Kocaeli, Turkey. *Marine Pollution Bulletin* 68: 134–139.

Blake, G.R. and Hartge, K.H. (1986). Bulk density. In: *Methods of Soil Analysis Part 1*:*Physical and Mineralogical Methods. (Edited by Klute, A. et al.)*, Monograph No. 9,Soil Science Society of America, Madison, Wisconsin, USA. pp. 363-375.

Böhner, J. and Selige, T. (2006). Spatial prediction of soil attributes using terrain analysis and climate regionalization. *Göttinger Geographische Abhandlungen* 115: 13-27.

Brocca, L., Tullo, T., Melone, F., Moramarco, T., Morbidelli, R. (2012). Catchment scale soil moisture spatial–temporal variability. *Journal of Hydrology* 422–423: 63–75.

Brus, D. J., Kempen, B., Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science* 62: 394 – 407.

Buchanan, B. P., Fleming, M.,Schneider, R. L.,Richards, B. K., Archibald, J., Qiu, Z.and Walter, M. T. (2014). Evaluating topographic wetness indices across central New Yorkagricultural landscapes. *Hydrological Earth Systems Science*18: 3279 – 3299.

Chen, D., Fu X.-Q., Wang, C., Liu X.-L., Li, H., Shen, J.-L., Wang, Y., Li, Y. Andwu, J.-S. (2015). Nitrous Oxide Emissions from a Masson Pine Forest Soil in Subtropical Central China. *Pedosphere* 25(2): 263–274.

Crommelinck, S., Bennett, R., Gerke, M., Nex, F., Yang,M. Y. and Vosselman, G. (2016). Review of Automatic Feature Extraction fromHigh-Resolution Optical Sensor Data for UAV-BasedCadastral Mapping. *Remote Sensing* 8 (689): 1-28.

DLR (2016a). New 3D world map – TanDEM-X global elevation model completed. [http://www.dlr.de/dlr/en/desktopdefault.aspx/tabid-10081/151_read-19509/#/gallery/24516.] site visited4/10/2016.

DLR (2016b). TanDEM-X Science Service System. [https://tandemx-science.dlr.de/] site visited 4/10/2016.

Gee, G. W. and Bauder, J. W. (1986). Particle size analysis. In: *Methods of Soil Analysis Part 1: Physical and Mineralogical Methods.* (Edited by Klute, A. et al.), Monograph 9, Soil Science Society of America, Madison, Wisconsin, USA. pp. 383 - 411.

Geological Survey of Tanzania (2016). Geological maps [http://www.gst.go.tz/mapproducts.html]. site visited 31/09/2016.

Gray, J. M., Bishop, T. F. A. and Yang, X. (2015a). Pragmatic models for the prediction and digital mapping of soil properties in eastern Australia. *Soil Research* 53: 24–42.

Gray, J. M., Bishop, T. F. A. and Wilson, B. R. (2015b). Factors Controlling Soil Organic Carbon Stocks with Depth in Eastern Australia. S*oil Science Society of America Journal*79:1741–1751.

Hengl, T., de Jesus J. M., MacMillan R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., Rosa, A. S., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Gonzalez, M. R. (2014). SoilGrids1km — Global Soil Information Based on Automated Mapping. *PLoS ONE* 9(8): e105992.

Hengl, T. (2016). GSIF: Global Soil Information Facilities. R package version 0.5-3. [https://CRAN.R-project.org/package=GSIF] site visited

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* 12(2): e0169748

Heuvelink, G. B. M. (2014). Uncertainty quantification of GlobalSoilMap products. In: *GlobalSoilMap: basis of the global spatial soil information system*. *Proceedings of 1st GlobalSoilMap Conference*(Edited by Arrouays, D.et al.).7-9 October 2013. Orleans, France. CRC Press. Leiden – Netherlands. pp. 335-340

Iezzoni A. F., Pritts M. P. (1991). Applicationsof principal components analysisto horticultural research. *Horticultural Science* 26: 334-338.

Jenny, H. (1941). *Factors of soil formation: A System of Quantitative Pedology*. 1st Edition (3rd Impression). McGraw Hill Book Company Inc, New York and London. 271pp.

Ji, W., Viscarra Rossel, R. A. and Shi, Z. (2015). Accounting for the effects of water and the environment on proximally sensed vis–NIR soil spectra and their calibrations. *European Journal of Soil Science* 66(3): 555–565.

Junior, W. de C., Chagas, C. da S., Lagacherie, P., Filho, B. C. and Bhering, S. B. (2014). Evaluation of statistical and geostatistical models of digital soil properties mapping in tropical mountain regions. *Revista Brasileira de Ciência do Solo*38: 706 - 717

Kaingo, J. and Tumbo, S. D. (2016). Soil Mapping and Web-GIS Development for Trans-Sec Project: Final Report. Sokoine Univesity of Agriculture, Morogoro - Tanzania. 88pp.

Kavianpoor, A., Ouri, A. E., Jafarian Jeloudar, Z., Kavian, A. (2012). Spatial Variability of Some Chemical and Physical Soil Properties in Nesho Mountainous Rangelands. *American Journal of Environmental Engineering* 2(1): 34 – 44.

Kempen, (2011). Updating soil information with digital soil mapping. Thesis submitted for award of PhD Degree of Wageningen University, Wagenigen - Netherlands. 218pp.

Kodešová, R. (2009). Soil micromorphology use for modeling of a non-equilibrium water and solute movement. *Plant Soil Environment* 55 (10): 424 – 428.

Kovačević, M., Bajat, B., Gajić, B. (2010). Soil type classification and estimation of soil properties using support vector machines. *Geoderma* 154: 340 - 347.

Kuhn, M., J. Wing, Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y. and Candan, C. (2016b). caret: Classification and Regression Training. R package version 6.0-64. [https://CRAN.R-project.org/package=caret] site visited on 31/10/2016

Kuhn, M., Weston, S., Keefer, C. and Coulter, N. (2016a). C code for Cubist by Ross Quinlan (2016). Cubist: Rule- And Instance-Based Regression Modeling. R package version 0.0.19. [https://CRAN.R-project.org/package=Cubist]site visited on 31/10/2016

Lark, R. M. (2012). Distinguishing spatially correlated random variation in soil from a ‘pure nugget’ process. *Geoderma* 185–186: 102–109.

Levi, M. R., Schaap, M. G. and Rasmussen, C. (2015). Application of Spatial Pedotransfer Functions to Understand Soil Modulation of Vegetation Response to Climate. *Vadose Zone Journal* 14 (9): 1-14.

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

Ließ, M., Schmidt, J., Glaser,B. (2016). Improving the Spatial Prediction of Soil Organic Carbon Stocks in a Complex Tropical Mountain Landscape by Methodological Specifications in Machine Learning Approaches. *PLoS ONE* 11(4): e0153673.

Liu, F., Zhang, G.-L., Sun, Y.-J., Zhao, Y.-G. and Li, D.-C. (2013). Mapping the Three-Dimensional Distribution of Soil Organic Matter across a Subtropical Hilly Landscape. *Soil Science Society of America Journal* 77: 1241 – 1253.

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

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A. (2014). Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method. *Geoderma* 235–236: 59–73.

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

Miháliková, M., Özyazıci, M. A. and Dengiz, O. (2016). Mapping Soil Water Retention on Agricultural Lands in Central and Eastern Parts of the Black Sea Region in Turkey. Journal of Irrigation and Drainage Engineering 142(12): 1-9.

Minasny, B., McBratney, A. B. and Walvoort, D. J. J. (2007). The variance quadtree algorithm: Use for spatial sampling design. *Computers & Geosciences* 33: 383–392.

Moore, I. D., Gessler, P. E., Nielsen, G. A. and Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* 57: 443 – 452.

Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., Martin, M. P., Arrouays, D. (2016). National versus global modelling the 3D distribution of soil organic carbon in mainland France. *Geoderma* 263: 16 – 34.

Nelson, D. W. and Sommers, L. E. (1982). Total carbon, Organic Carbon, and Organic Matter. In: *Methods of Soil Analysis. Part 2 - Chemical and Mineralogical Properties.* (Edited by Page, A. L. et al.), Monograph 9. American Society of Agronomy, Madison, Wisconsin, USA. pp. 539 - 579.

Orton, T. G., Pringle, M.J., Bishop, T.F.A. (2016). A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. *Geoderma* 262: 174.–.186.

QGIS Development Team, (2016). QGIS Geographic Information System. Open Source Geospatial Foundation Project. [http://www.qgis.org/] site visited 31/09/2016.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [http://www.R-project.org/] site visited 31/09/2016.

SAS (2017). Principal Component Analysis. [http://support.sas.com/publishing/pubcat/chaps/55129.pdf] site visited 15/11/2017

Stoorvogel, J.J., Kempen, B., Heuvelink, G.B.M., de Bruin, S. (2009). Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. *Geoderma* 149: 161.–.170.

Trans-Sec (2017). Trans-SEC - Innovating Strategies to safeguard Food Security using Technology and Knowledge Transfer: A people-centred Approach. http://www.trans-sec.org/ site visited 20/03/2017

USGS (2017). SRTM 1 Arc-Second Global. [https://earthexplorer.usgs.gov/] site visited 31/01/2017.

Ustuner, M., Sanli F. B. and Dixon, B. (2015) Application of Support Vector Machines for Landuse Classification Using High-Resolution RapidEye Images: A Sensitivity Analysis. *European Journal of Remote Sensing* 48: 403.-.422.

Veronesi, F., Corstanje, R. and Mayr, T. (2012). Mapping soil compaction in 3D with depth functions. *Soil & Tillage Research* 124:111 – 118.

Viloria, J. A., Viloria-Botello, A., Pineda, M. C., Valera, A.(2015). Digital modelling of landscape and soil in a mountainous region: A neuro-fuzzy approach. *Geomorphology* 253: 199–207.

Viscarra-Rossel, R. A., Adamchuk, V. I., Sudduth, K. A., McKenzie, N. J. and Lobsey, C. (2011). Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. In: *Advances in Agronomy*

(Edited by Sparks, D. L.) Vol. 113, Academic Press, Burlington. pp. 237-282.

VSN International (VSNi) (2012). GenStat for Windows 15th Edition. VSN International, Hemel Hempstead, UK. https://www.vsni.co.uk/software/genstat/ site visited February 2017

Wang, X., Cai, Q., Ye, L., Qu, X. (2012). Evaluation of spatial and temporal variation in stream water quality by multivariate statistical techniques: A case study of the Xiangxi River basin, China. Quaternary International 282: 137 – 144.

World Reference Base (WRB) (2015). *World reference base for soil resources 2014, update 2015: International soil classification system for naming soils and creating legends for soil maps.* World Soil Resources Reports No. 106. FAO, Rome. 127pp.

Young and Pearce, (2013). A Beginner’s Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology* 9(2): 79 - 94.

CHAPTER FIVE

5.0 CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSIONS

The following conclusions were drawn from the study:

i. Stratified random sampling design has slightly better accuracy than the spatial coverage sampling design. The stratified random sampling design is a probabilistic sampling design and also allows for inferential statistical measures like confidence limits to be drawn from the soil moisture holding capacity data during mapping.

ii. Support vector regression algorithmexhibits good accuracy for generating pedo-transfer functions in comparison to multiple linear regression approach. Apparently, usinglinear kernel based support vector machines offers primaryimprovements for observed accuracy.However, SVR-PTFs are not a straight forward implementation.

iii. Machine learning algorithms and soil depth functions had relatively good accuracy for 3-dimensional mapping of soil moisture holding capacity.Random forests machine learning algorithm had better accuracy than cubist algorithm. Spatial prediction patterns of the soil moisture holding capacity maps follow the patterns of spatial covariate layers used. Prediction accuracy appears to be influenced by the resolution of the spatial covariate layers. There is a slight reduction in prediction accuracy with a combination of measured and estimated moisture holding capacity data.

5.2 RECOMMENDATIONS

The following recommendations are proposed:

i. Soil sampling as implemented in this study can be ideally applied for any digital soil mapping task where fresh soil sampling has to be performed. However, in the circumstances where some data already exists at known spatial locations, spatial infilling of the less sampled areas can be performed.

ii. Developed SVR-PTFs are best suited as a backend implementationwith a graphical user interface in a modelling framework like a soil information system, crop or hydrological models. Therefore, further work will be needed towards that end. However, the multiple-linear regression PTFs can easily be implemented as an alternative means. The utility of the SVR-PTFs might be limited to soil conditions of the dry sub-humid climates and there will be need to further explore and extend its robustness for application across other pedo-environments. The greatest hindrance though is the limited availability of soil moisture data. It will therefore be of essence to collate available soil moisture data to build a soilshydraulic properties database for Tanzania to facilitate further soil hydrological studies.

iii. Implementation of the 3-D mapping considers the influence of spatial covariates as constant across all depths. This may not be consistentas the spatial covariates are often representative of surface features. Therefore, further research to better account for variability or influence of spatial covariates with depth. Also, how the prediction accuracy of the machine learning technique could be improved using finer resolution spatial covariates from proximal sensing and unmanned aerial vehicles needs to be studied.

iv. In the 3-D predictive mapping approach adopted, a 50 % even split of data was performed to assess the influence of substitution of measured data with PTF-estimated data on mapping accuracy. This split was arbitrary.It will therefore be important to explore the desirableproportion of sampling points to substitute with PTF-estimated data for ‘optimal’ accuracy. This will provide a better understanding of the trade-off for sampling costs and prediction error for mapping soil moisture holding capacity.

APPENDICES

Appendix 1: Field data of representative pedon for *Kichanga*

Profile number: P1-KCH/ILA-MIH

Authors: Jacob Kaingo and Mwango Date: 190215 Weather: SU/WC4

Region: Morogoro District: Kilosa Ward: Ulaya Village: Ilakala

Description of Location: About 150 m West of Ilakala Catholic Church

Coordinates: 7° 7'45.15"S/36°55'49.91"E Elevation: 549 m asl.

Indigenous Mapping Unit: Kichanga WRB Soil Name:Hyperdystric Cambisol (loamic, ochric)

Soil Temperature Regime: Isohyperthermic Soil Moisture Regime: Ustic
Parent material: Quartzite

Landform: Undulating. Slope: Gently sloping North, Straight Convex, Upper-Slope. LandUse: Traditional Rainfed arable cultivation. Human Influence: Ploughing, Raised beds and Bunding. Surface characteristics: None. Erosion: None. Deposition: None. Natural drainage class: Well drained.

Ap 0 - 16/28 cm: Brown (7.5 YR 4/4) moist; sandy loam; friable moist, non-sticky and non-plastic wet; weak fine and medium crumbly structure; many medium and coarse pores, few fine pores; many fine roots and common medium roots; abrupt wavy boundary

AB 16/28 - 60 cm: Brown (7.5 YR 4/4) dry, dull brown (7.5YR 6/3) moist; sandy loam; slightly hard dry, friable moist, non-sticky and non-plastic wet; moderate fine to coarse subangular blocks; very few medium and coarse roots, common fine roots; few burrows present; gradual smooth boundary

Bw1 60 - 85 cm: bright brown (7.5 YR 5/6) dry, dull orange (7.5 YR 6/4) moist; sandy loam; slightly hard dry, friable moist, slightly sticky and slightly plastic wet; moderate fine to coarse subangular blocks; common fine, few medium, and very few coarse roots; few burrows present; gradual smooth boundary

Bw2 85 - 150/160 cm: orange (7.5 YR 6/8) dry, bright brown (7.5 YR 5/8) moist; sandy loam; soft dry, friable moist, sticky and plastic wet; weak to moderate fine and medium subangular blocks; very few medium, very few coarse, and few fine roots; clear wavy boundary

C 150/160 - 200+ cm: Gravelly layer with numerous quartzite residual rocks.

Appendix 2: Field data of representative pedon for *Ngunja*

Profile number: P2-NGJ/ILA-MIE

Authors: Jacob Kaingo and Mwango Date: 200215 Weather: PU/WC4

Region: Morogoro District: Kilosa Ward: Ulaya Village: Ilakala

Description of Location: About 500 m South West from Nyaranda Primary School

Coordinates: 7° 8'17.95"S/36°54'57.01"E Elevation: 610 m asl.

Indigenous Mapping Unit: Ngunja WRB Soil Name: Rhodic Acrisol (clayic, cutanic, epieutric, profondic)

Soil Temperature Regime: Isohyperthermic Soil Moisture Regime: Ustic

Parent material:

Landform: Plateau. Slope: Very gently sloping North, convex, summit. LandUse: Rainfed arable cultivation. Human Influence: Ploughing, Raised beds and Bunding. Surface characteristics: None. Erosion: Water, slight sheet and rill, active at present. Deposition: none. Natural drainage class: Well drained

Ap 0 - 18/24 cm: Very dark reddish brown (2.5 YR 2/4) moist; sandy clay; friable moist, slightly sticky and slightly plastic wet; weak fine subangular blocks; many medium and few fine pores; common fine, and few medium roots; abrupt wavy boundary

AB 18/24 – 40/50 cm: dark red brown (2.5YR3/6) dry, dark red brown (2.5YR 3/4) moist; clay; slightly hard dry, friable moist, sticky and slightly plastic wet; moderate medium and coarse subangular blocks; many medium and few fine pores; many fine, few medium, and very few coarse roots; termites and few burrows present; clear wavy boundary

Bt1 40/50 - 100 cm: red (10 R 4/6) dry, dark red (10R 3/4) moist; clay; hard dry, friable moist, sticky and slightly plastic wet; moderate to strong fine, medium, and coarse subangular blocks; many medium and few fine pores; few fine, very few medium, and very few coarse roots; few burrows present; diffuse smooth boundary

Bt2 100 - 145/160 cm: red (10 R 4/6) dry, dark red (10R 3/4) moist; clay; slightly hard dry, friable moist, sticky and slightly plastic wet; moderate fine and medium subangular blocks; thin clay-iron cutans present; many medium and few fine pores; very few fine roots; few burrows present; diffuse wavy boundary

Bt3 145/160 – 200+ cm: red (10 R 4/8) dry, dark red (10R 3/6) moist; clay; slightly hard dry; friable moist, sticky and slightly plastic wet; moderate fine and medium subangular blocks; thin clay-iron cutans present; many medium and few fine pores; few burrows present

Appendix 3: Field data of representative pedon for *Tifu-Tifu*

Profile number: P3-TFT/ILA-SHU

Authors: Jacob Kaingo Date: 210215 Weather: PU/WC3

Region: Morogoro District: Kilosa Ward: Ulaya Village: Ilakala

Description of Location: About 800 m east off Kilosa-Mikumi main road at Ilakala Mosque

Coordinates: 7° 8'10.50"S/36°56'18.76"E Elevation: 549 m asl.

Indigenous Mapping Unit: Tifu-Tifu WRB Soil Name: Luvic Stagnic Umbrisol (endoeutric, loamic)

Soil Temperature Regime: Isohyperthermic Soil Moisture Regime: Ustic

Parent material: Quartzite

Landform: Depression. Slope: Gently sloping South, Straight, convex, Back Slope. LandUse: Rainfed arable cultivation. Human Influence: Ploughing, Raised beds and Bunding. Surface characteristics: None. Erosion: Water, slight sheet and rill, Active at present. Deposition: none. Natural drainage class: Well drained

Ap **0 - 30 cm**: very dark grey (7.5 YR 3/1) moist; sandy loam; friable moist, slightly sticky and slightly plastic wet; moderate fine and medium crumbly structure; common fine and medium pores; common fine, common very fine, and very few medium roots; termites present; gradual smooth boundary to

AB 30 – 50/60cm: very dark grey (7.5 YR 3/1) moist; sandyloam; friable to firm moist, slightly sticky and slightly plastic to plastic wet; moderate to strong fine to coarse lumpy structure; common fine and medium pores; very few very fine, few fine, and common medium roots; termites present; very few medium distinct clear yellowish brown mottles; very few soft fine elongated nodules; gradual wavy boundary to

Bgw 50/60 - 75/81 cm: dark brown (7.5YR3/2) moist; sandy loam; firm to very firm moist, sticky to very sticky and plastic to very plastic wet; strong coarse subangular blocky structure; common fine and medium pores; very few very fine, and very few fine roots; many fine distinct clear yellowish brown mottles; very few fine hard round concretions; clear smooth boundary to

Btg 75/81 - 98/112 cm: olive yellow (2.5Y 6/8) moist; sandy clay loam; soft dry, friable moist, sticky to very sticky and plastic to very plastic wet; strong coarse and very coarse subangular blocky structure; broken thin clay cutans present; few fine hard round concretions; very few fine hard irregular residual rock fragments; common fine and very fine, few medium pores; very few fine, and few medium roots; abundant coarse prominent diffuse yellowish brown mottles; abrupt irregular boundary to

Bgc 98/112 - 122/134 cm: Pale brown (2.5Y 7/4) moist; sandy clay loam; very firm moist, slightly sticky to sticky and slightly plastic wet; strong very coarse/thick subangular and prismatic structure; common fine hard round concretions; abundant coarse prominent diffuse yellow mottlesclear wavy boudary to

C 122/134 - 200+ cm: Gravelly layer with dominant fine hard round red-black concretions and abundant medium and coarse hard angular residual rock fragments.

Appendix 4: Field data of representative pedon for *Wakitope-Mweusi*

Profile number: P4-MWS/ILA-JUU

Authors: Jacob Kaingo Date: 210215 Weather: SU/WC3

Region: Morogoro District: Kilosa Ward: Ulaya Village: Ilakala

Description of Location: About 2.2 km South-west from Makondeko sub-village trading centre

Coordinates: 7° 8'36.38"S/36°52'16.43"E Elevation: 625 m asl.

Mapping Unit: Wakitope Mweusi WRB Soil Name: Endogleyic Protovertic Eutric Cambisol (colluvic, ruptic)

Soil Temperature Regime: Isohyperthermic Soil Moisture Regime: Udic

Parent material: Colluvial deposits of quartzite and metamorphic feldspar rocks. Landform: Valley floor. Slope: Nearly Level North; Straight, valley bottom. LandUse: Rainfed arable cultivation. Human Influence: Ploughing. Surface characteristics: None. Erosion: None. Deposition: Water deposition, Active in recent past. Natural drainage class: Well drained

Ap 0 - 22/34 cm: black (10 YR 2/1) moist; clay loam; very few medium distinct sharp very pale brown (10 YR 7/4) mottles; friable moist; sticky and slightly plastic to plastic wet; weak to moderate very fine to medium crumbly structure; common fine and very fine pores; common fine, many very fine, very few medium roots; clear wavy boundary to

2AB 22/34 – 42/49 cm: dark brown (10 YR 3/3) moist; sandy clay loam; common fine faint clear red mottles (.5 YR 4/8); very friable moist, slightly sticky and slightly plastic to plastic wet; weak very fine to medium granular and crumbly structure; few fine and very fine pores; many medium and coarse fresh or slightly weathered angular quartz gravel, few medium weathered sub-rounded feldspar gravel; few fine, common very fine, and very few coarse roots; clear wavy boundary to

3Bw 42/49 – 60/64 cm: yellowish red (5YR 4/6) moist; sandy; very friable moist, non-sticky and non-plastic wet; weak fine and very fine single grain structure; medium fine and pores; abundant medium and coarse fresh or slightly weathered angular quartz gravel; very few fine roots, common very fine roots; clear wavy boundary to

4Bhw 60/64 - 80/93 cm: very dark brown (7.5YR 2.5/3) moist; clay loam; very few fine distinct clear reddish yellow mottles (5 YR 6/8); firm to very firm moist, very sticky and very plastic wet; strong coarse subangular blocks; many fine and very fine pores; very few fine and common very fine roots; very few fine roots; clear irregular boundary to

5Bhg 80/93 – 102/106 cm: dusky red (2.5YR 3/2) moist; clay; common fine distinct clear red mottles (2.5 YR 4/6); extremely firm moist, very sticky and plastic wet; strong coarse and very coarse prismatic structure; many very fine pores; very few fine weathered sub-rounded feldspar gravels; very few very fine, and very few fine roots; few burrows present; clear broken boundary to

6Abh 80/106 – 121/141 cm: black (10YR 2/1) moist; clay loam; very firm moist, sticky to very sticky and slightly plastic to plastic wet; strong coarse and very coarse prismatic structure; common fine and very fine pores; Very few fine strongly weathered sub-rounded quartz gravels; very few very fine roots, few fine roots, and very few coarse roots; clear irregular boundary to

7Bwg 121/141 - 150/160 cm: dark brown (7.5 YR 3/4) moist; clay; many medium prominent diffuse dark red mottles (2.5 YR 3/6); extremely firm moist, very sticky and plastic to very plastic wet; strong coarse and very coarse prismatic structure; common fine pores; very few fine fresh or slightly weathered angular quartz gravel, few fine fresh or slightly weathered sub-rounded feldspar gravel;very few fine roots; gradual wavy boundary to

8Blc 150/160 – 200+ cm: dark red brown (5YR 3/4) moist; sandy clay loam; abundant medium distinct diffuse very dark bluish gray mottles (10 B 3/1); extremely firm moist, slightly sticky and slightly plastic wet; strong coarse and very coarse prismatic structure; many fine pores; abundant fine and medium fresh or slightly weathered sub-rounded and rounded feldspar gravels; very few fine roots; numerous red-black concretions

Appendix 5: Field data of representative pedon for *Mfinyanzi*

Profile number: P6-MFZ/ILA-MAS

Authors: Jacob Kaingo Date: 230215 Weather: PC/WC3

Region: Morogoro District: Kilosa Ward: Ulaya Village: Ilakala

Description of Location: Mashineni sub village, about 500 m east from Kilosa to Mikumi Road

Coordinates: 7° 9'7.62"S/36°56'18.02"E Elevation: 564 m asl.

Mapping Unit: Mfinyanzi WRB Soil Name: Pellic Vertisol (ferric, humic, mesotrophic)

Soil Temperature Regime: Isohyperthermic Soil Moisture Regime: Ustic

Parent material:

Landform: Depression. Slope: Very gently sloping East; Straight, lower slope. Surface characteristics: None, Erosion: WS and WR, Active at present. Deposition: none. Natural drainage class: Well drained

Ap 0 - 40 cm: black (10 YR 2/1) moist; clay; very firm moist, very sticky and very plastic wet; strong coarse to very coarse subangular blocks; many very fine pores; many very fine roots, many medium roots; burrows; diffuse smooth boundary to

AB 40 – 64/71 cm: very dark grey (10 YR 3/1) moist; clay; very firm moist, very sticky and very plastic wet; strong coarse to very coarse subangular blocks; many very fine pores; common fine roots, common very fine roots; burrows; very few fine hard and soft round reddish to yellowish red ferruginous concretions; clear wavy boundary to

Bw1 64/71 - 130 cm: dark brown (7.5 YR 3/3) moist; sandy clay; very firm moist, sticky to very sticky and plastic to very plastic wet; strong coarse to very coarse prismatic structure; burrows, termites present; many fine and medium pores; few very fine roots, very few fine roots; common very fine both hard and soft irregular white siliceous soft segregations, common medium both hard and soft round red ferruginous concretions; few fine and medium freshly or slightly weathered angular quartz gravel; termite nests present; gradual smooth boundary to

Bw2 130 - 175 cm: dark brown (7.5 YR 3/2) moist; sandy clay; very firm moist, sticky and plastic wet; strong coarse to very coarse prismatic structure; common fine and medium pores; few very fine and very few fine roots; few medium both hard and soft round reddish to yellowish red ferruginous concretions, common very fine hard and soft irregular white siliceous soft segregation; termite nests present; gradual smooth boundary;

Bw3 175 – 200+ cm: very dark gray (7.5 YR 3/1) moist; clay; very firm moist, very sticky and very plastic wet; strong coarse to very coarse prismatic structure; medium fine and very fine pores; very few very fine, and very few fine roots; very few very fine hard and soft irregular white siliceous soft segregation

Appendix 6: Attributes of lithological classes of Geology layer

CODE	FORMATION	TIME	SOILS	ROCK_TYPE
MBG	Usagarani System - Basic - Ultrabasic Intrusiva	Precambrian	Meta gabbroic rocks, meta pyroxienite, metaperitite	Meta - Igneous Rocks
NTS	River Alluvium	Post Miocene	Orange-red soils and undifferentiated soils on precambrian rocks. Sandy to stony loams and clays	Unconsolidated superficial deposits and soils
RAL	River Alluvium	Quaternary	Clays, loams, sands, gravels	Unconsolidated superficial deposits and soils
XAB	Usagarani system-acid gneisses (magmatitic in part)	Precambrian	Biotite gneiss	Meta - Sedimentary Rocks