Densification of spatially-sparse legacy soil data at a national scale: a digital mapping approach

Stephen Itolo Chibuzo Akpa

B.Agriculture (Hons)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Environmental Science
Faculty of Agriculture and Environment
The University of Sydney
New South Wales
Australia

2016

Certificate of originality

I hereby declare that the intellectual content of this submission is the product of my own work and to the best of my knowledge it contains no material previously published or written by another person, or material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where reference is made in the text. Any contribution made to the research by others, with whom I have worked, in the project's design and conception or in style, presentation and linguistic expression is explicitly acknowledged.

Signature: Stephen Akpa

Date: 31/08/2015

Statement of Authorship

Contribution of co-authors

Chapter 3

Akpa, S. I.C., I.O.A. Odeh, T.F.A. Bishop & A.E. Hartemink, 2014. Digital mapping of soil particle-size fractions for Nigeria. *Soil Science Society of America Journal*, 78, 1953-1966.

- Associate Professor Inakwu O. A. Odeh supervised this work, contributed to discussions, and edited the manuscript.
- Associate Professor Thomas F. A. Bishop contributed to discussions, and edited the manuscript.
- Professor Alfred E. Hartemink contributed to discussions, and edited the manuscript.

Chapter 4

Akpa, S. I.C., I.O.A. Odeh, T.F.A. Bishop, A.E. Hartemink, I.Y. Amapu, 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma*, 271, 202-215.

- Associate Professor Inakwu O. A. Odeh supervised this work, contributed to discussions, and edited the manuscript.
- Associate Professor Thomas F. A. Bishop contributed to discussions, and edited the manuscript.
- Professor Alfred E. Hartemink contributed to discussions, and edited the manuscript.
- Professor Ishaku Y. Amapu contributed to discussions, and edited the manuscript.

Chapter 5

Akpa, S. I.C., I.O.A. Odeh, T.F.A. Bishop, S.U. Ugbaje, 2016. Enhancing pedotransfer functions with environmental attributes for estimating selected soil properties in a data-scarce situation. *Soil Use and Management* – Revised.

- Associate Professor Inakwu O. A. Odeh supervised this work, contributed to discussions, and edited the manuscript.

- Associate Professor Thomas F. A. Bishop contributed to discussions, and edited the manuscript.
- Mr Sabastine U. Ugbaje contributed to data analysis, and edited the manuscript.

Chapter 6

Akpa, S. I.C., I.O.A. Odeh, T.F.A. Bishop, K.O. Oluwasemire, 2016. Multi-criteria evaluation of irrigation suitability at national scale: Application of Choquet Integral. *International Journal of Digital Earth*-Submitted.

- Associate Professor Inakwu O. A. Odeh supervised this work, contributed to discussions, and edited the manuscript.
- Associate Professor Thomas F. A. Bishop contributed to discussions, and edited the manuscript.
- Dr Kolawole O. Oluwasemiri edited the manuscript and provided useful comments.

Contribution of senior author to all the papers

- Conceptualize and design the research
- Carried out the data analysis
- Drafted all the manuscripts
- Submitted the manuscript to the journal and addressed the reviewer's comments.

I hereby certify that the above statement about my contribution to the research work in this PhD thesis is true and accurate, and I give Stephen Akpa full permission to submit this work as part of his PhD thesis.

Name	**Signature**	**Date**
Inakwu O.A Odeh		
Thomas F.A Bishop		
Alfred E. Hartemink		
Ishaku Y. Amapu		
Kolawole O. Oluwasemiri		
Sabastine U. Ugbaje		

Thesis Summary

Spatially continuous and quantitative soil information is an integral component of most management decisions on agriculture and the environment. This is of particular importance in resource-poor countries, especially in sub-Saharan Africa (SSA) mostly plagued by poverty, hunger and land degradation. However, quantitative soil information is not readily available in the right format. Digital soil mapping (DSM) has been a viable approach to providing spatial soil information but its adoption in most resource-poor countries, especially at the national scale is limited by inadequate or low spread of data. Therefore, the focus of this thesis is on developing and/or optimizing existing DSM techniques for the densification of spatially-sparse legacy soil data at the national scale. The specific research objectives include to: (a) determine the appropriate prediction model for digital mapping of key soil properties at a national scale, (b) estimate total carbon and carbon sequestration potential of soils at a broad scale, (c) develop and/or calibrate pedotransfer functions appropriate in a data-scarce situation and (d) assess irrigation suitability for national agricultural planning using DSM products. In executing the set objectives, legacy soil data for Nigeria was utilized as the main dataset for model building and application.

First, the robustness of Random Forest model (RFM) was tested in predicting soil particle-size fractions (PSFs) using legacy soil data and covariates. To improve PSFs prediction, soil sampling depth was introduced as predictor variable while additive log-ratio (ALR) transformation technique was applied to ensure that predicted PSFs some up to a constant value of 100. Results indicated good prediction accuracy with RFM while the inclusion of sampling depth as a predictor substantially improved prediction accuracy, especially at the lower depth intervals. Nigerian soils are predominantly coarse-textured especially in the northern region of the country. Soil texture ranges from sand (4.2×10^6 ha) to sandy loam (5.3×10^7ha) in the surface layers and from sandy clay loam (5.2×10^7 ha) to clay (6.9×10^6 ha) in the subsoils.

Second, in order to quantify the carbon sequestration capacity of soils, soil organic carbon (SOC) and bulk density (BD) were predicted using legacy soil data from which SOC density and stock were calculated. SOC density was then overlaid with land use land cover

(LULC), agro-ecological zone (AEZ) and predominant soil maps to quantify the carbon sequestration of soils and their variation across and within the different AEZs. Results showed that about 6.5 Pg C with an average density of 71.60 Mg C ha^{-1} abound in the top 1 m soil depth while soils in the Derived and Sahel Savannahs have the largest capacity to sequester additional carbon.

Furthermore, to improve the performance of BD and exchangeable cation exchange capacity (ECEC) pedotransfer functions (PTFs), the combination of soil and environmental data was explored. Input datasets were first divided into topsoils and subsoils according to soil horizon depth while MLR and RFM were then fitted to estimate BD and ECEC respectively. Results showed that subdividing the input data based on soil depth significantly improved the accuracy of PTFs in estimating BD and ECEC. However, the combination soil and environmental data only improves BD estimation. Important predictors of BD include sand, silt, elevation, rainfall and temperature for estimation at topsoil while EVI, elevation, temperature and clay are the most important BD predictors in the subsoil. Also, clay, sand, pH, rainfall and SOC are the most important predictors of ECEC in the topsoil while pH, sand, clay, temperature and rainfall are the most important predictors of ECEC in the subsoil.

Finally, to support informed decision making and national agricultural planning, the application of Choquet fuzzy integral (CI) aggregation technique in irrigation suitability assessment was assessed. This was achieved through multi-criteria analysis of potential evaluation criteria including soil, climatic and landscape attributes. Results indicate that CI is a better aggregation operator compared to the classical weighted mean. A total of 3.34×10^6 ha (approximately 4% of total land area) is suitable for surface irrigation in Nigeria while major limitations are due to topographic and soil attributes. Also, majority of current irrigation projects are situated in moderate to marginally irrrigation suitable areas.

In conclusion, this work has revealed how relatively sparse national soil database can be populated to support decisions on national agricultural and environmental planning. The thesis examined appropriate DSM techniques applicable to sparse-data conditions. This is the first systematic study on operational DSM at the national scale especially in the Sub-

Saharan Africa (SSA). The findings of this research will provide quantitative basis for framing appropriate policies on sustainable food production and environmental management in the SSA and other resource-poor countries of the world.

Acknowledgements

My profound gratitude goes to God Almighty who gave me the strength, health and ability to accomplish this work. To Him alone be glory, honour and adoration forever, Amen. This work would not be possible without the support of people who influenced and encouraged me along the way. I wish to express my sincere appreciation to my principal supervisor, Associate Professor Inakwu Odeh, whose works captured my interest in the field of pedometrics and digital soil mapping way back in 2008. Thank you for your educative and prompt supervision, guidance, useful suggestions, constructive and objective criticisms and constant encouragement, without which this work would have been a mere dream. Special thanks to my associate supervisor, Associate Professor Thomas Bishop for his kind supervision, valuable comments and technical supports. I feel utmost pride in acknowledging the guidance and support received from Professors Ishaku Amapu and Joshua Ogunwole prior to the commencement and throughout the period of my study.

I am deeply indebted to the University of Sydney for providing the University of Sydney International Scholarship (USydIS) during my study period. Thanks also to the Ahmadu Bello University for granting me fellowship to undertake this study. I wish to acknowledge the efforts of the soil scientists in Nigeria who carried out various soil surveys, which were the original sources of data used for this work. I extend my sincere appreciation to the management of ISRIC in the Netherlands and other institutions as custodians of the soil survey legacy data used in this study. My profound gratitude also goes to Professor Alfred Hartemink of the University of Wisconsin, USA for his invaluable comments and technical supports during earlier publications that form part of this thesis. The entire thesis work would not be possible, without the assistance of individuals like Dr W.B. Malgwi, Dr G. A. Ajiboye, Dr K.O. Oluwasemire, Dr G.E. Akinbola, Dr. R. C. Eneje, Dr B.T. Udoh, Dr E.U. Onweremadu, Prof. B.D. Tarfa, Prof. S.S. Noma, Prof. V.O. Chude, Prof. T.O. Ibia, Prof. G. Lekwa, Prof. F.O.R. Akamigbo, Prof. C.A. Asadu, and Prof. F.N. Ikpe. These people facilitated access to data in their respective institutions during my fieldtrip to Nigeria.

I also wish to appreciate the supports and companionship of past and present colleagues at the Faculty of Agriculture and Environment, especially the Hydrology and GIS Laboratory

(HGISL) Research Group: Dr. Senani Karunaratne, Dr Ana Horta, Dr Richard Muita, Dr Nirmala Liyanage, Dr Jessica Health, Dr Kanika Singh, Dr Brendan Malone, Dr Nathan Odgers, Dr Uta Stockmann, Dr Vicky Verts, Dr Mana Gharun, Phil Hughes, Jose Padarian, Tshering Dorji, Ronald Muchelo, Mulatua Hailu, Mary Otieno, Cheryl Poon, Dipankar Kundu, Farzina Akter and many others that this space will not permit me to mention. I appreciate the positive impacts of staff members of the Faculty of Agriculture and Environment, especially members of my annual progress review interview panels in the first two years of my candidature: Professor Budiman Minasny, Associate Professors Stephen Cattle, Willems Vervoort and Brett Whelan, for their advice, criticisms, supports and encouragements. Many thanks to Drs Peter Thompson, Peter Amp, Daniel Tan, Floris, Mrs Annette Vervoort, Ms Pamela and Ms Fiona Lawrence, for their assistance at one point or the other during my candidature at the faculty.

Special thanks also to my inestimable friends that kept supporting and encouraging me all the way, especially Sabastine Ugbaje, Jenny Adaobi, Achikwu Manasseh, Chigozie Ezegbe, and Farzane Haghighi. I am deeply indebted to the families of Edward Winters, Odeh Inakwu (who doubled as my supervisor), Seyi Lagoke, Zechariah Martings, David George and Pastor John Adeyemi for making my stay in Sydney a seamless and memorable one. Many thanks to the Winters’ for accommodating my family and for helping out during our relocation at the tail end of this work. Special thanks to Retmun Zechariah for her help with data entry after my fieldtrip to Nigeria. I sincerely appreciate the love and kindness shown to me by Ms Virginia Madu of the Department of Soil Science, AB.U Zaria. She’s always on the lookout for my family and prayed fervently for the success of this work.

Words alone are inadequate to express my profound gratitude to my affectionate parents, who have always wished to see me soaring high in the skies of success, and inspired me immensely. Many thanks also to my siblings, cousins, uncles, aunties and in-laws for their unrivalled love, moral support and fervent prayers. I show a deep appreciation and regards for the kindness, affection and support of my heartthrob whose encouragements, prayers and supports made it possible for me to accomplish this research work. Dear, you are indeed God’s special gift and favour to my life and I will forever be grateful. To my little

Angel and Princess Chisom, thanks for your love and understanding (right from your conception until birth) without which I would not have met my thesis submission deadline.

Finally, I wish to appreciate the support and fellowship of the entire parishioners of Our Lady of Lourdes Parish Earlwood, All Saints Parish Liverpool, the RCCG Convent House Alexandra, the Partaker's Sanctuary Blacktown and the Igbo Catholic Community, Harris Park. The success of this research is indeed an answer to your collective prayers and goodwills.

Blessed be God forever, Amen.

Dedication

In loving memory of my late grandmother, Mary Eyum Onyilokwu, who laid the foundation of my academic prowess and taught me the most beautiful lessons in life. May the Almighty God grant her soul an eternal repose in the bosom of Abraham. Amen.

Publications

Papers published or submitted for publication in peer review scientific journals:

Akpa, S.I.C., I.O.A. Odeh, T.F.A. Bishop & A.E. Hartemink, 2014. Digital mapping of soil particle-size fractions for Nigeria. *Soil Science Society of America Journal*, 78, 1953-1966. **(Chapter 3)**

Akpa, S.I.C., I.O.A. Odeh, T.F.A. Bishop, A.E. Hartemink & I.Y. Amapu, 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma*, 271, 202-215. **(Chapter 4)**

Akpa, S.I.C., I.O.A. Odeh, T.F.A. Bishop & S.U. Ugbaje, 2016. Enhancing pedotransfer functions with environmental attributes for estimating selected soil properties in a data-scarce situation. *Soil Use and Management* – Revised. **(Chapter 5)**

Akpa, S.I.C., I.O.A. Odeh, T.F.A. Bishop, K.O. Oluwasemire, 2016. Multi-criteria evaluation of irrigation suitability at national scale: Application of Choquet Integral. *International Journal of Digital Earth*-Submitted. **(Chapter 6)**

Other publications and presentations made from this thesis:

Abstracts

Akpa, S., Odeh, I, Bishop, T. 2013. Assessing the potential carbon sequesterability of soils under different land use types across agro-climatic zones of Nigeria. *10th Biennial Meeting of Commission 1.5 Pedometrics Division 1 of the International Union of Soil Science (IUSS)*, ICRAF & CIAT. Kenya August 26-September 2, 2013.

Akpa, S., Odeh, I., & Bishop, T. 2014. Operational digital soil mapping at national scale: Application of Random Forest to spatial prediction of soil particle-size fractions. *20th World Congress of Soil Science*. Jeju, South Korea June 8-13, 2014.

Posters

Akpa, S.I.C., Odeh, I.O.A, Bishop, T.F.A. 2014. Utilization of Nigeria legacy soil data for digital mapping of functional soil properties. *The Annual Faculty Symposium, Faculty of Agriculture and Environment, The University of Sydney, Australia* July 15, 2014.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Amapu, I.Y., Sanni, B.M., Ajiboye G.A. 2015. Digital soil mapping for national irrigation: Integration of DSM products with other biophysical data for irrigation assessment. *11th Biennial Meeting of Commission 1.5 Pedometrics Division 1 of the International Union of Soil Science (IUSS)*, The University of Cordoba, Spain September 14-18, 2015.

Contents

List of Figures

List of Tables

Chapter 1.

General introduction, scope of study and research objectives

1.1 Background

Soil resources are integral part of human existence and need to be carefully managed to ensure their security and sustainable development (Arrouays et al., 2014; McBratney et al., 2014). As such, knowledge of soil is prerequisite to meaningful management decisions on food production and environmental quality. Soil information is required for any viable soil-related agro-technology transfer, environmental modelling and monitoring, urban planning and developmental policies. However, soil information is not readily available in the required format (Greve et al 2012). One promising approach to quick delivery of required soil information is the use of quantitative soil modelling techniques such as digital soil mapping (Hartemink and McBratney, 2008).

Digital soil mapping (DSM) is a timely, reliable and cost effective way to acquiring continuous soil information. Basically, DSM involves establishing a relationship between a target soil attribute (sampled at sparse locations) and the so-called *scorpan* factors or environmental variables (Minasny et al., 2013; Minasny and McBratney, 2016). Prediction is thereafter made at unobserved (dense) locations using the grids of readily available environmental variables at those locations along with the associated prediction error (Nelson et al., 2011). The interpolation of soil properties at unvisited location is made possible through the use of advanced numerical models, most of which are spatially explicit and data-driven. However, in many instances, there is often not enough soil data to form statistical structure for high resolution spatial prediction using these models. This has hindered the practicality of DSM in many countries of the world.

As spatially explicit soil data are integral part of DSM operations the success of any DSM campaigns will depend on available funds or investments for additional soil survey and sampling (Hartemink et al., 2013). While this poses little or no problem in most rich or developed countries like the United States, Australia and countries of Europe, it is a major hindrance in many data-poor countries. Even more worrisome is the fact that there is a greater and more pressing need for quantitative soil information in these data-poor countries than in their developed counterparts. In addition, there is no likelihood of future investments on soil surveying and mapping in most of these countries (Cambule et al., 2015). Therefore, considering the prohibitive cost requirements of any new soil survey scheme, legacy soil data will be an integral part of most DSM operations especially in the developing parts of the world.

Fortunately, many countries in the developing parts of the world, especially in the sub-Saharan Africa (SSA) are covered by legacy soil surveys (Odeh et al., 2012). However, most of these soil surveys differ in spatial scale, coverage, objectives, age and quality. As such, legacy data obtained from these surveys have inherent peculiarities and uncertainties that may negate their use in DSM. These include uneven data spread, missing and/ or incomplete data, and mixture of data types, among others. The challenge therefore is on how to transform this useful but often inadequate soil data to support national developmental planning. To achieve this will require the development of new and/or optimization of existing DSM techniques to make them suitable for legacy soil data especially in sparse data conditions.

The basis of this study is to explore the use of modern DSM techniques for the densification of sparse legacy soil data to support national scale developmental planning. It provides the first comprehensive study on operational digital soil mapping at national scale with a broader application in the developing parts of the world, especially in the SSA. The study was undertaken using legacy soil data from Nigeria. Nigeria which is the most populous country in SSA has a rich history of soil survey (Odeh et al., 2012). Despite boosting an estimated 71.2 million hectares of arable land (Ayoola, 2009); Nigeria is currently faced with a daunting challenge of ensuring national food security as well as combating land degradation and desert encroachment. Tackling these challenges will require adequate information on the soil resources to support sustainable agricultural intensification and environmental management.

1.2 The purpose of this study

Specifically, the objectives of this study include:

1. To determine the appropriate prediction model for digital mapping of key soil properties at a national scale.
2. To estimate total carbon stock and sequestration potential of soils at a broad scale.
3. To develop and/or calibrate pedotransfer functions appropriate for a data-scarce situation.
4. To assess irrigation suitability for national agricultural planning using digital soil mapping products.

1.3 The scope of study

The organization of this thesis is such that each chapter is focused on tackling an aspect of value-addition to sparse legacy soil data to meeting modern needs for soil information. Chapter 2 is a literature review, covering such issues as the need for DSM and soil information, especially in the developing country, the challenges of DSM in the absence of dense soil samples as well as the pressing need for legacy soil data. It further highlighted the challenges posed by legacy soil data and discussed different DSM techniques that may be relevant under sparse data conditions.

Following on from the literature review, Chapter 3 deals with prediction of particle-size fractions as a compositional data using random forest model (RFM). Then Chapter 4 covers a crucial aspect of national environmental planning; estimation of carbon stock and sequestration potential of soils under different land use and agro-ecological zones. In Chapter 5, two different techniques in the form of combination of soil and environmental data and input data grouping based on soil depth were evaluated for their enhancement of pedotransfer functions (PTFs)'s performances on sparse datasets for bulk desity and effective cation exchange capaciry estimation at national scale. In Chapter 6 the robustness of Choquet integral multi-criteria aggregation techniques was tested for irrigation suitability assessment using some of the soil and environmental data derived in Chapter 3. Finally, Chapter 7 presents the key research findings from this thesis and a pointer to future research.

The research chapters (Chapters 3-6) presented here are formatted as either published articles or submitted manuscript. Therefore, some sections covering background information in these chapters appear to be unavoidably overlapped; these are not deliberate repetitions. Also, the reference style of the first paper (Chapter 3) has been adopted throughout the thesis to ensure consistency.

References

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J. 2014. Chapter Three-GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. *Advances in Agronomy,* 125, 93-134.

Ayoola, J.B. 2009. National medium term priority framework for food security and agricultural development in Nigeria. FAO, Rome,Italy.

Cambule, A., Rossiter, D., Stoorvogel, J., Smaling, E. 2015. Rescue and renewal of legacy soil resource inventories: A case study of the Limpopo National Park, Mozambique. *Catena*, 125, 169-182.

Greve, M.H., Bou Kheir, R., Greve, M.B., P.K. Bøcher. P.K. 2012. Using digital elevation models as an environmental predictor for soil clay contents. *Soil Science. Society of American Journal*, 76: 2116–2127. doi:10.2136/sssaj2010.0354

Grunwald, S., Vasques, G.M., Rivero, R.G. 2015. Chapter One-Fusion of Soil and Remote Sensing Data to Model Soil Properties. *Advances in Agronomy* 131, 1-109.

Hartemink, A.E., Krasilnikov, P., Bockheim, J. 2013. Soil maps of the world. *Geoderma,* 207, 256-267.

Hartemink, A.E., McBratney, A. 2008. A soil science renaissance. *Geoderma*, 148, 123-129.

McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. *Geoderma,* 213, 203-213.

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

Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I. 2013. Digital mapping of soil carbon. *Advances in Agronomy*, 118, 1-47.

Nelson, M.A., Bishop, T.F.A., Triantafilis, J., Odeh, I.O.A., 2011. An error budget for different sources of error in digital soil mapping. *European Journal of Soil Science*, 62, 417-430.

Odeh, I.O., Leenaars, J., Hartemink, A., Amapu, I. 2012. The challenges of collating legacy data for digital mapping of Nigerian soils. In: Minasny, B., Malone, B.P., McBratney, A. (Eds). Digital Soil Assessments and Beyond. Taylor & Francis Group, London, pp 453-458.

Chapter 2.

Digital soil mapping under sparse legacy data situation: a review of literatures.

2.1 Introduction

There are numerous inherent challenges associated with the use of legacy soil data for digital soil mapping operations (Minasny et al., 2013). These include sparse distribution of data points, differing quality and data age, different sampling schemes and analytical techniques. Therefore, this chapter is aimed to provide general insights into the various DSM approaches suitable for predicting the spatial distribution of soil attributes using legacy soil data. Firstly, the needs for soil information, particularly in the SSA were highlighted. Thereafter, the underlying principle of DSM and the various techniques at the disposal of digital soil mapping scientists were reviewed with the purpose of elucidating appriopriate models applicable to legacy soil data.

In addressing the challenge of sparse distribution or limited data point associated with legacy data, this chapter provides an overview of PTFs as well as techniques used in improving the performance of PTFs. Finally, in a way of utilizing DSM products for decision-making on national developmental projects, the multi-criteria decision-making approaches were reviewed in the light of finding more robust approach to land suitability assessment. Such suitability assessment could provide the basis for sound management decision-making in the implementation of better soil management strategies to support sustainable agricultural intensification especially in the developing countries.

2.2 The need for soil information in developing countries

As shown in Fig. 2.1 below, soils provide many fundamental ecosystem goods and services such as food security, biodiversity protection, climate change adaptation and environmental regulation (Grunwald et al., 2015). As such, there has been a renewed interest in detailed, accurate and spatially continuous soil information for the purpose of management decisions regarding these goods and services (Arrouays et al., 2014). Soil information is required for developing recommendations on best management practices to improve soil productivity, increase crop yield through selection of appropriate crop variety and irrigation water needs of crops (Omuto et al., 2013; Grealish et al., 2015). Unfortunately, the existing soil databases in most countries are neither exhaustive nor accurate enough for promoting a credible use of the soil information for the purpose of these management decisions especially at the national scale. This is particularly overwhelming in most resource-poor countries, particularly in the SSA.

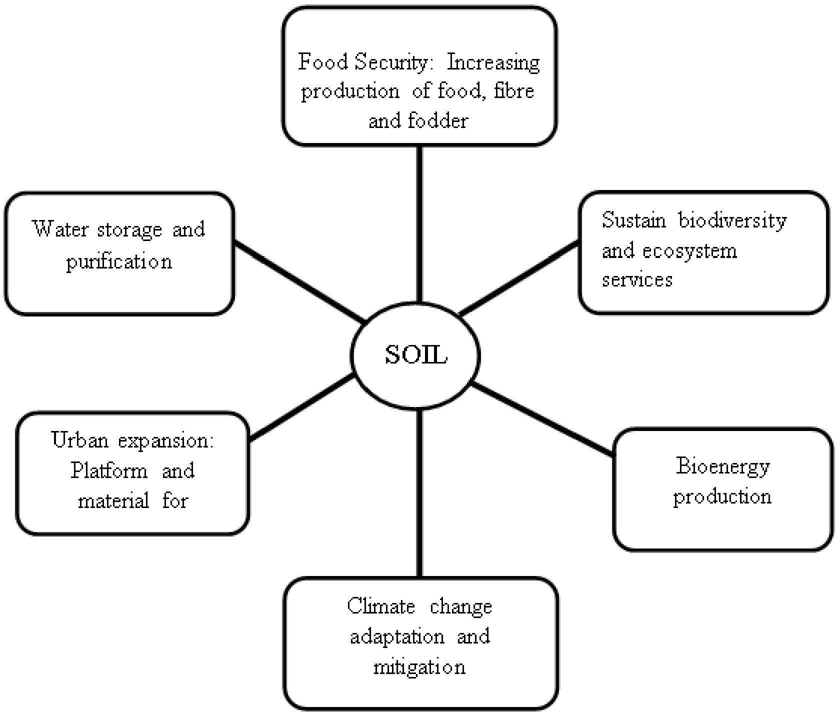


Figure 2.1 Importance and interaction of soil with human needs.

Adapted and modified from Omuto et al 2013.

In SSA, agriculture is the dominant economic activity, providing source of livelihood to about 60-80 % of the teeming population. According to FAO (2009) agricultural production in SSA is below optimum production level and will need to increase significantly to meet the food needs of its population that is expected to double by 2050. However, only about 20 % of the expected production increase may arise from possible land expansion (FAO, 2009). This is in sharp contrast to previous reports that the increase in agricultural production in the SSA in the last 50 years was achieved through expansion of arable lands. Therefore, meeting up with the future demands from agriculture in the SSA will require well-informed decisions on selection of agricultural lands through strategic development of areas with more productive soils. In the context of this strategic selection of agricultural lands, soil mapping will play a vital role in planning and implementation of enterprises to increase agricultural production in the SSA (Omuto et al., 2013). Additionally, soil information is needed for impact assessments of current land use and management on soil functioning (Lal, 2008; Northcliff, 2009) so as to combat land degradation problems that is prevalent in most part of the region.

2.3 The need for digital soil mapping in sub-Saharan Africa.

As mentioned in the previous section, soil information is very important for sustainable agricultural development in most countries of the SSA. Identifying the importance of soil information, particularly in the context of food and fibre production and sustainable development (Hartemink et al., 2013), soil surveys at various spatial extents and scales have been carried out in SSA (see Table 2.1). These surveys were predominantly carried out using traditional soil mapping approaches and in the instances of past colonial masters and/or foreign developmental aid organizations (Odeh et al., 2012). A typical example of such survey efforts include the USDA initiated and assisted national soil inventory project for Nigeria that led to the production of soil map of the country (at the scale of 1:650,000) in the late 80s (FDALR, 1990; Odeh et al., 2012). Another example is the Rwanda’s conventional national soil survey (at a scale of 1:50,000) carried out between 1981 and 1994 (Van Ranst et al., 2010).

Generally, these previous soil survey efforts vary in their objectives and as such were not designed to cover large extents using statistical sampling schemes. Thus, they are not representative of the overall condition of soils in SSA. These surveys however, provide a wealth of legacy soil data and information which can be enhanced for effective guide in agricultural land use development. The challenge however, is that in their current state these legacy soil data cannot adequately meet modern demands for quantitative soil information. Therefore there is need to transform them to meet modern soil information needs through quantitative soil-landscape modelling approach such as DSM.

Despite the obvious need for DSM in SSA, the adoption of DSM in this region is still at the juvenile stage especially for national scale applications. DSM efforts made so far in the SSA are largely through international aid support and individual contributions. Examples include the digital soil map of Africa (Hengl et al., 2015), digital SOC map of Southeastern Kenya (Alejandra Mora-Vallejo et al., 2008), digital SOC map of the Senegalese Peanut Basin (Stoorvogel, et al., 2009) and more recently the SOC map of Limpopo national park in Mozambique (Cambule et al., 2014). Among these studies only Hengl et al. (2015) at continental scale and Cambule et al. (2014) at the field scale have utilized legacy soil data. To meet the increased soil information demand in sub-Saharan Africa, a regional project called Africa Soil Information Service (AfSIS) was launched in 2009 under the guidance of the GlobalSoilMap.Net consortium (Sanchez et al., 2009). One of the key activities of AfSIS is to set standards for soil data collection and soil evaluation in the SSA and to coordinate the gathering of legacy soil data for DSM activities in the region.

Table 2.1 National soil survey coverage in 21 African countries (Modified from Van Ranst et al., 2010).

Country	Small scale 1:500,000 - ± 100,000 (%)	Medium scale 1:100,000 - ± 50,000 (%)	Large scale <1:25,000 (%)
Algeria	-	5	5
Benin	100	10	2
Botwana	40	5	-
Burkina Faso	100	25	-
Burundi	100	-	-
Cameroon	30	5	1
DR Congo	10	5	-
Egypt	100	10	10
Gabon	30	-	-
Gambia	100	-	100
Ghana	95	-	-
Kenya	100	25	-
Mali	50	-	-
Morocco	-	40	20
Nigeria	70	35	0.6
Rwanda	100	100	-
South Africa	70	-	-
Swaziland	100	10	5
Tanzania	50	-	-
Togo	80	20	-
Uganda	100	-	1

2.4 Data requirements for digital soil mapping

Data required for digital soil mapping include geo-referenced soil data and environmental variables or covariates (Minansy et al., 2008). The geo-referenced soil data are used to establish the variation of soil attributes across a particular landscape while the covariates are required to support the application of predictions of that variation across the entire area of interest (MacMillan, 2008). Soil data could be sourced from existing soil maps, auger point observation or through the use of proximal and remote sensing techniques. However, due to time and budget constraints, legacy soil data could be the most widely used soil data for DSM

(Minasny et al., 2013). Depending on the purpose of the DSM operation, soil data could be observations on soil bulk density, pH, soil organic carbon (SOC), effective cation exchange capacity (ECEC), particle size fractions (clay, silt, sand), water holding capacity, and hydraulic conductivity, among others (Arrouays et al., 2014). Generally, the density of soil data points geographically and the resolution of the outputs of DSM are determined by the extent of the mapping area, the purpose of required soil information and availability of adequate budget.

A key approach to enhancing the use of legacy soil data in DSM, especially those in data-sparse countries, is the use of covariates. Past studies suggested that the most popular covariates include geology or lithology, digital elevation model and its derivatives such as slope gradient, aspect, curvatures, compound topographical index, quasi-dynamic wetness index, stream power index, multi-resolution index of valley bottom flatness (MRVBF), etc. Increasingly, other covariates include the gamma radiometric products (K, U, Th), electromagnetic induction (EM) data, multi-temporal satellite images or bands and their derivatives such as enhanced vegetation index (EVI), normalized difference vegetation index (NDVI), soil-adjusted vegetation index, among others (McBratney et al., 2000; 2003; Minansy et al., 2008, Schmidt et al., 2014). Although there are no standard thresholds for optimum number of covariates in DSM, their intensity depends on the target soil attribute, the choice of prediction model used as well as the pedogenesis of the area to be mapped (Yang et al., 2011). On a general note, covariates for DSM are chosen on the basis of data availability and the researcher's expert knowledge (Miller et al., 2015). According to Minansy et al. (2013), terrain attributes and remote sensing data in their various forms are the most widely used covariates in DSM operation.

2.5 The challenges of using legacy soil data for DSM

As alluded to above, in many countries of the world particularly in the SSA, legacy soil data is the primary source of data available for DSM. However, the use of legacy soil data can be very challenging due to inherent inadequacies in them. Legacy soil data are products of traditional soil survey that were mostly carried out without proper statistical sampling design. As such, there are inherent biases in legacy data so much that they may not be representative of the geographical landscape. One of such biases is the issue of low density or uneven spread of data. This sparse coverage of quantitative observations could introduce considerable spatial uncertainty. This is of particular concern with soil attributes such as SOC that are highly dynamic over short range (Powers et al., 2011). In Africa, reduced number of data has been reported to limit the use of most reconnaissance maps (Mora-Vallejo et al., 2008; Stoorvogel et al., 2009). Another inherent bias in legacy soil data is the issue of varying data age. Legacy soil data are collected at different times and for different purposes thereby resulting in data with wide differences in currency. This could limit their use for soil monitoring purposes.

Locational inaccuracy of legacy soil data poses another challenge in DSM. The inaccuracy may be due to the fact that the pre-1990 soil surveys were carried out in pre-GPS era and as such were not properly geo-referenced. Improperly geo-referenced input data in DSM could increase positional error which is usually transmitted to the overall uncertainty of the predictive maps. This could be due to data points being assigned to the wrong covariate values (Grimm and Behrens, 2010). Other limitations of using legacy soil data for DSM include missing or incomplete information, mixture of both categorical and numerical data as well as varying soil layer (horizon) interval and soil profile depth. In using legacy data for DSM, it is therefore important to understand these limitations. Some knowledge of the purpose and methods of the soil surveys as sources of the legacy data could indicate the quality and any bias in the spread of the samples.

2.6 DSM techniques suitable to sparse legacy soil data condition.

2.6.1 Harmonization of varying legacy soil profile depth interval

In traditional soil survey, soil samples are often collected by genetic horizon with the assumption that the horizon value of a given soil attribute represents its mean value for the depth interval of that horizon (Odgers et al., 2012). However, in environmental modelling, most models require soil information at specified depth ranges rather than the pedogenetic horizons (Adhikari et al., 2013). In a typical legacy soil database, it is not uncommon to have different survey reports with data for different combinations of horizons and depths. To circumvent the challenge posed by varying horizon depth, DSM techniques using continuous depth functions or splines have been developed to map soil properties at specified depth intervals (Bishop et al., 1999; Malone et al., 2009). A mass-preserving or equal area quadratic spline consists of a series of quadratic polynomials that join at the "knots" located at the horizon boundaries (Bishop et al., 1999). It passes through each soil horizon, and thus maintains the average value of the soil attributes. The "knots" are linear between horizons but quadratic within the horizons; giving a linear-quadratic smoothing spline. Detailed background knowledge of spline has been elaborated by Bishop et al. (1999) and Malone et al. (2009). However, for the sake of brevity, a summary of the mass-preserving spline algorithm following Malone et al. (2009) is provided here.

For a given soil profile and a given soil property, the boundaries of the n horizons are denoted by $x_0 < x_1, ... < x_n$. The soil property values, y_i (i = 1…n) could be modelled mathematically as:

$$y_i = \bar{f_i} + e_i \qquad [2.1]$$

where $\bar{f_i}$ is the mean value of $f(x)$ over the interval (x_{i-1}, x), and e_i represents the measurement error with mean 0 and variance σ^2. Finally $f(x)$ is the spline function, which is found by minimizing:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{f}_i\right)^2 + \lambda\int_{x_0}^{x_n}\acute{f}\ (x^2)dx \qquad [2.2]$$

The first term in Eq.2.2 represents the fit of the spline to the data while the second term is the roughness of the function$f(x)$. The lambda (λ) controls the trade-off between the fit and the roughness of the spline. Several (λ) values have been tested for various soil attributes but the value of 0.1 has been reported to give good results for a number of soil attributes (Bishop et al., 1999, Odgers et al., 2012; Adhikari et al., 2013). Spline function are limited in their capacity to estimate soil attribute under abrupt changes in the soil properties, especially in the case of interpolating the particle-size fractions of texture contrast or duplex soils and change from topsoil to subsoil OC values of peat soils. In such conditions it has been recommended to introduce a quasi or pseudo horizon to the existing profile data (Odgers et al., 2012; Adhikari et al., 2013).

2.6.2 The use of appropriate prediction model

DSM employs several modelling techniques for spatial prediction of soil class and attributes. These techniques can be broadly grouped into two major categories (i) spatial prediction models including geostatistical models (e.g. kigring), statistical models (e.g multiple linear regression) and their hybrids (e.g. regression kriging), and (ii) data mining tools such as regression or decision trees, neural networks, boosting machines and fuzzy systems. Generally the spatial prediction models are suitable for data-rich situations while the data mining tools or machine learning models are suitable for sparse data and/or complex situations (Hastie et al., 2009). Since there have been a fair bit of seminar review works on most of these DSM models (McBratney et al., 2000, McBratney et al., 2003, Scull et al., 2003), the focus of discussions here is on selected techniques that are suitable to sparse legacy data condition.

2.6.2.1 Data mining tools or machine learning models

There is an increasing use of data-mining or machine learning prediction techniques for spatial soil prediction (Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2015; Talaab et al., 2015a). Data-mining tools were designed to explore patterns in complex data and generate models fitted with many parameters (Hastie et al., 2009). In general, their strength is the ability to use continuous and categorical predictors, and the fact that they are very robust relative to predictor specifications. These techniques are capable of catering for relatively complex structure in legacy soil data that may be difficult to detect with many conventional geostatistical tools. In addition, data mining models do not require the *a priori* specifications of a model to relate explanatory with dependent variables, but rather use an algorithm to learn the form of those relationships (Breiman, 2001). In the following sub-sections, an overview of some of the data mining models that could accommodate the inherent complexity in legacy soil data especially in sparse conditions is provided.

2.6.2.1.1 Classification and regression tree

Originally developed in the early 1980s, the classification and regression tree (CART) algorithm (Breiman et al. 1984) was first applied to predictive soil mapping in the early 90s (Lagacherie, 1992). Since then various studies have highlighted the efficiency of CART for spatial prediction of soil properties at various scales (McKenzie and Ryan 1999; Moran and Bui, 2002; Henderson et al. 2005; Scull et al., 2005; Barthold et al., 2008; Vasques et al., 2008; Stoorvogel et al., 2009). One of the most interesting features of CART for DSM is that it gives quantitative insight into the input data using explicit splitting rules. In addition, it can uncover relatively important predictor variables by counting the times the variables were used in the tree nodes (Bui et al., 2006). CART has several advantages over classical linear regression models that make them a better suit for legacy soil data: it is non-sensitive to missing data, perform automatic variable subset selection, and can handle both quantitative and categorical data. However, CART has been criticized for overfitting in model derivation, especially in the presence of noise or outliers (Lagacherie and Holmes, 1997; McKenzie and Ryan 1999).

2.6.2.1.2 Artificial neural networks

In contrast to CART, artificial neural networks (ANNs) are non-parametric data mining tools which are analogous to neural networks of the human brain (Venables and Ripley, 1994). ANNs can be used to model complex relationships between inputs and outputs or to find hidden patterns in a given data set (Tveito, 2010). One important feature of ANNs is their adaptive nature through "learning" during the classification or prediction process. This makes them a powerful and popular modelling technique for solving complex and non-linear processes. ANNs can achieve linearization of the predicted outputs by weighting the network inputs with non-linear sigmoid or logistic functions and summing them to derive the non-linear response. In spite of these advantages, ANNs are criticized as being "black-box" models and require higher computational power than most prediction models. As part of DSM techniques, the ANNs have been predominantly used for predicting soil class and deriving pedotransfer functions (Minasny et al., 2002; Botula et al., 2015).

2.6.2.1.3 Boosted regression trees

Boosted regression trees (BRT) belong to the gradient boosting modelling (GBM) family of statistical algorithms (Collard et al., 2014). They employ CART approaches to make prediction of a target variable. However, BRT improve prediction accuracy compared with CART by minimizing the risk of over-fitting and thus improves prediction power (Schapire et al., 1998; Lawrence et al., 2004). Boosting techniques are generally applied to increase performance of a given estimation method by generating instances of the method iteratively from a training data set and additively combining them in a forward "stagewise" procedure (Elith et al., 2008). Like most data mining prediction models BRT has the inherent ability to represent interactions among predictor variables without a priori knowledge of their distribution. Additionally, BRT is resistant to the effects of outliers, missing data and autocorrelation among variables (Jalabert et al., 2010) and more importantly like CART, it can work with both qualitative and quantitative variables (Friedman and Meulman, 2003). Two main parameters are required for the fitting of BRT: the learning rate and the tree size or interaction depth. BRT has been employed in soil science for soil organic carbon prediction (Martin et al., 2011; 2014; Collard, et al., 2014).

2.6.2.1.4 Random Forest

The Random Forest (RF) was developed as an extension of RTM to improve its prediction accuracy (Breiman 2001; Liaw and Wiener, 2002) and like BRT, to reduce model overfitting. RF is an assemblage of a number of classification or regression trees using two levels of randomization for every tree in the forest (Breiman, 2001). RF has several advantages over other prediction models: i) insensitivity to noise or weak prediction variables as it selects the most important variable at each node split (Okun and Priisalu, 2007), and ii) reasonable predictive performance with noisy predictive variables (Diaz-Uriarte and de Andres, 2006). In addition, RF trees are insensitive to missing values or outliers in a given dataset (Craig and Huettmann, 2008), a common feature with most legacy soil data. RF's major strength lies in its two randomization procedures of bootrapping and random input selection (Sequeira et al., 2014) and subsequent bagging of the predictions. RF has been vastly employed in remote sensing studies (Gislason et al., 2006; Lawrence et al., 2006) with substantial usage in ecology (Peters et al., 2008; Prasad et al., 2006) and genetics (Wu et al., 2009). However, there is a dearth of information on its application in soil science studies (Grimm et al., 2008, Viscara Rosel and Brehens, 2010). Several studies have demonstrated the superiority of RF to commonly available geostatistical and data mining prediction models in environmental research (Prasad et al., 2006; Li and Heap, 2008). In Soil Science, Ließ et al. (2012) reported a better performance of RF models than CART in predicting soil texture of the surface horizon.

2.6.2.1.5 Bayesian network models

Bayesian networks (BNs) are graphical probabilistic models developed in the 1980s from the branch of mathematics known as probabilistic reasoning (Peal, 1988). BNs apply probabilities derived from either measured data or expert opinion in making predictions. They present cause-effect relationships (one event leading to another) through several connections in a system of networks (Hough et al., 2010) and differ from common network based methods, such as ANNs, by allowing the integration of qualitative experts knowledge into the model structure. In this context, soil experts are allowed to judge whether the fitted model makes some pedogenic senses (Taalab et al., 2015a). BNs have several advantages over other regularly used modelling techniques in DSM (Taalab et al., 2015a; 2015b Brungard et al., 2015). Unlike purely deterministic models, BNs offer a structured method of handling the uncertainty associated with soil predictions by expressing the existing relationships between soil attributes or class and the covariates as a probability function (Dlamini, 2010).

Another major appeal of BNs is that, the integration of experts' knowledge in the model structure can be used to either supplement measured data or solely define soil-landscape relationships (Finke, 2012). This is an ideal way of addressing problems of limited data availability (Kuhnert et al., 2010, Kuhnert, 2011). BNs have been optimally applied to environmental studies such as ecology and natural resource management (McCann et al., 2006; Kuhnert et al., 2010), landscape conservation (McCloskey et al., 2011), habitat mapping (Smith et al., 2007), erosion risk mapping (Aalders et al., 2011) and wildfire risk mapping (Dlamini, 2010). However Bayesian modelling approaches have only been recently used in soil mapping (Mayr et al., 2010; Brungard et al., 2015; Lorenzetti et al., 2015; Taalab et al., 2015a; 2015b; Xiong et al., 2015; Yang et al., 2015).

2.6.2.2 Regression kriging

Regression kriging (RK) is a hybrid prediction technique that combines a regression (either simple or multiple-linear) of the target soil attribute on covariates with ordinary, or simple, kriging of the regression residuals (Odeh et al., 1995; Goovaerts, 1997; Hengl et al., 2007). In RK the assumption is that the deterministic component of the target soil attribute is accounted for through the regression model, while the model residuals represent the spatially varying but dependent component. Several variants of RK have been proposed and used in different studies (Odeh et al., 1994; Odeh et al., 1995; Knotters et al., 1995; Hengl et al., 2007) with slight modifications depending on the task at hand.

Regression-kriging is increasingly popular because it achieves lower prediction errors at the control points and to the relative availability of less espensive covariates. Several studies have demonstrated the superiority of regression kriging over other methods of interpolation such as ordinary kriging, universal kriging, multiple-linear regression and cokriging especially in soil studies (Odeh et al., 1995; Odeh and McBratney, 2000; Hengl et al., 2007; Li and Heap, 2008). Recently, it has been demonstrated that combination of machine learning methods like random forest with OK using RK approach can improve prediction accuracy significantly (Li et al., 2011). One major limitation of regression kriging is that the way in which the explanatory variables appear in the trend is highly empirical and thus may not reflect the actual physical processes (Odeh and McBratney, 2000). There is also the high computational demand as the analyst will have to carry various steps in different software within statistical and GIS environments (Hengl et al., 2007).

2.6.2.3 Fuzzy expert systems

In addition to the aforementioned data mining and hybrid models, other quantitative modelling techniques that have been employed in the use of legacy soil data for DSM are those based on the fuzzy set theory or fuzzy expert systems (Zhu et al., 2001). Fuzzy set theory is a generalization of the traditional set theory in that it modified the classical concept of belonging to a set to allow partial degrees of membership of a set i.e any values within a continuum range of 0 and 1 (McBratney and Odeh, 1997; Torbert et al., 2008). Fuzzy set theory is also a generalization of Boolean algebra more suitable to situations where there are zones of gradual transition compared to the conventional crisp boundaries used in dividing classes (Burrough et al., 1992). Originally formulated by Zadeh (1965), the fuzzy set theory builds on the traditional crisp two-valued theory of binary membership functions of TRUE or FALSE by adding intermediate values or partial membership. It is a mathematical method of quantifying ambiguity and vagueness such that data that do not have sharply defined boundaries are grouped into membership classes. Fuzzy expert systems have been applied in a variety of studies such as in remote sensing (Wang, 1990), soil pollution management (Amini et al., 2005), salinity study (Malins and Metternicht, 2006), soil classification (Odeh et al., 1992; Zhu et al., 1996; Qi et al., 2006) and land evaluation (Burrough et al., 1992; Davidson et al., 1994; Braimoh et al, 2004).

The development of fuzzy logic-based digital soil mapping techniques is due to its ability to represent the continuous nature of soil spatial variation (Zhu et al., 2001; Yang et al., 2005). Generally, in fuzzy expert system approaches, soil spatial parameters are expressed in terms of membership functions of different soil classes (McBratney et al., 2000). This is then used to produce conventional soil class maps or to forecast spatial parameters of specific soil properties (Zhu et al., 1996). Lagacherie (2005) proposed a procedure based on fuzzy pattern matching to translate soil class descriptions in soil database into a set of membership functions. Qi et al. (2006) developed a fuzzy soil mapping approach to represent soil-environment knowledge as fuzzy membership functions. Later Qi et al. (2008) developed a data mining method using the Expectation Maximization (EM) algorithm to define membership functions based on the information extracted from conventional soil class maps. One advantage of fuzzy approach to DSM is its low data requirement.

Fuzzy expert systems can sometimes be employed in conjunction with geostatistical spatial prediction techniques such as kriging to produce fuzzy soil maps of continuous classes (McBratney and De Gruijter, 1992; Odeh et al., 1992). The underlying process of fuzzy soil mapping involves using the fuzzy k-means classifier to the classify the soil surface and kriging the matrix of membership grades into a continuous soil surface classes which represents the different soil mapping units (SMU). There are no rigid boundaries demarcating these SMUs or geographical soil entities (Scull et al., 2003) hence, any individual location can belong to more than one class.

2.7 Pedotransfer functions for DSM of functional properties at national scale

Pedotransfer functions (PTFs) are viable alternatives to bridging the gap between available soil data and required data and could complement DSM efforts especially in developing countries (Minasny and Hartemink, 2011). PTFs are"predictive functions of certain soil properties derived from other easily measured properties" (Bouma, 1989). As such, the main focus of PTFs is to estimate, relatively difficult to measure functional properties (e.g. soil water holding capacity (SWHC), hydraulic conductivity (Ks), soil erodibility index, bulk density and pH-buffering capacity) from primary soil properties (e.g. particle-size fractions, organic carbon, and pH). The estimated functional soil properties are highly desirable but often not available in most national soil databases.

2.7.1 Development of pedotransfer functions

One of the most commonly used modelling techniques in the development of PTFs is multiple linear regression (MLR). In MLR, all readily available predictor variables are linearly related to the target soil data (Abbasi et al., 2011) to estimate the target data. In addition to the MLR, more complex techniques have also been developed as PTFs. These include artificial neural networks (Minasny and McBratney, 2002; Merdun et al., 2006), support vector machines (Lamorski et al., 2008; Twarakavi et al., 2009; Jafarzadeh et al., 2015), genetic programming (Padarian et al., 2014), group method of data handling (Nemes et al., 2005; Ungaro et al., 2005; Nemes and Rawls, 2006., Vereecken et al., 2010) and nonparametric nearest neighbor (Nemes et al., 2006; 2010). Other relatively new approaches include boosted regression trees (Martin et al.,2009; Jalabert et al.,2010; Ghehi et al., 2012, Jordan et al., 2015),) and random forest (Sequeira et al., 2014). Several studies have employed one or more of these techniques to develop PTFs with varying level of performance.

Minansy et al. (1999) compared both parametric and point estimates of water retention curves using PTFs developed from multiple linear regression (MLR), extended nonlinear regression (ENR) and artificial neural network (ANN). They reported that ENR out performed MLR and ANN in parametric PTF prediction. However, when the number of input parameters is greater than three, ANNs usually perform better than regression techniques, particularly under low uncertainty conditions (Baker and Ellison, 2008; Minansy et al., 2004). Lake et al. (2009), in a different study, also reported the superiority of ANN models over MLR models which they attributed to the ability of ANN to establish a non-linear relationship between the dependent and independent variables. Botula et al. (2015) also reported a better performance of kNN PTFs over MLR. In contrast, Merdun et al. (2006) reported no significance differences between the accuracy of MLR and ANN models in point estimate of soil hydraulic conductivity. However, they opined that MLR predicted point and parametric variables of soil hydraulic parameters are intuitively better than those of ANN.

2.7.2 Limitations of pedotransfer functions for digital soil mapping

As discussed previously, several studies in the soil science and hydrology community have been involved in developing PTFs from available soil databases around the world. Nonetheless, the reliability of many of these PTFs is dependent on the size and structure of the input data (Romano and Chirico, 2004; Haghverdi et al., 2012). For instance, in a relatively small area, with low spatial soil variability and homogenous terrain, high reliability could be obtained from a reasonably few number of soil samples (Ghehi et al., 2012). However, in a given large and heterogeneous landscape characterized by high soil spatial variability, reliability of PTFs may be impacted by the size and spread of the soil samplings.

Another major limitation of PTFs in DSM is their requirement of independent observations and no spatial autocorrelation (Brus and de Gruijter, 1997). Consequently, there is need to carefully evaluate the class domain of new datasets with the aim to calibrate the datasets before any attempt to extrapolate PTFs beyond their original statistical training limits and geographical area (Ungaro et al., 2005). For example, Medina et al. (2002) reported that water retention PTFs developed for soils in the USA and Europe cannot be used for Ferralsols in Cuba. Bell and van Keulen (1996) found that field capacity data from disturbed soil samples overestimates in-situ field capacity for all soils except for coarser textured soil. Hence they cautioned the use of field capacity data derived from disturbed samples.

In contrast to the above opinions on transferability of PTFs, Cresswell et al. (2006) reported a good transferability of soil water retention capacity PTFs developed from Australia soil data to French soils. Similarly, Manyame et al. (2007) also found that Campbell's PTFs for water retention and hydraulic conductivity function could be applied for sandy soils in Niger but with a rather modest accuracy. As such there is no universal validity of any particular PTF (Bastet et al., 1999); therefore new PTFs should be validated with new data sets in the domain of the calibration datasets (Rab et al., 2011). To facilitate PTFs validation on new datasets, McBratney et al. (2011) recommended that three tables containing information and statistics of the calibration data, predicted variables and the validation data should accompany any published PTFs to enhance effective usage.

2.7.3 Enhancing the performance of pedotransfer functions for scarce data condition

The reliability or performance of PTFs is dependent on the size and structure of the input data which is a major concern in many cases. However, to improve the performance of PTFs, several techniques have been employed prior to fitting PTFs with varying level of success. These include stratification of measured data based on soil taxonomy (Manrique & Jones, 1991; Heuscher et al., 2005) or by soil horizons and the incorporation of additional variables such as soil physiographic and morphological properties such as soil consistence and structure (Calhoun et al., 2001), horizon designation (Jalabert et al., 2010), etc. The concern however, is that most of these soil morphological properties (e.g. soil consistence and structure), are not always available from soil survey data (Manrique & Jones, 1991; Calhoun et al., 2001; Heuscher et al., 2005).

Other studies have reported improvement of PTFS following the incorporation of environmental data such as topography and vegetation attributes to primary soil properties (Pachepsky et al., 2001; Leij et al., 2004; Sharma et al., 2006; Jana and Mohanty, 2011; Wang et al., 2014). For instance, Pachepsky et al. (2001) used a combination of different topographical attributes and soil physical data to develop PTFs in predicting soil hydraulic properties for hill-slope soils in the USA using linear regression models. Their results showed significant improvement in the performance of PTFs in predicting soil hydraulic properties. Leij et al. (2004) also corroborated their reports in a similar study conducted in Italy. Similarly, Sharma et al. (2006) reported the incorporation of different combinations of topographic, vegetative and soil attributes into PTFs as reliable methods to estimating soil moisture contents. Recently, Wang et al. (2014) used the combination of soil basic properties and terrain attributes to develop a PTF for estimating bulk density (BD) across the Loess Plateau in China. They concluded that the addition of slope gradient to soil physical properties could estimate BD with reasonable accuracy.

2.8 Prediction of Soil particle size fraction as a compositional data

The relative soil composition of sand, silt and clay fractions (particle size fraction) which determines the texture is inarguably the most important soil physical property that controls most physical, chemical and biological processes in the soil (Adhikari et al., 2013, Safari et al., 2013). For instance the particle-size distribution of soil can greatly influence the soil water retention capacity (Botula et al., 2012), plant nutrient retention capacity (Kettler et al., 2001), leaching and erosion potential of soils (Thompson et al., 2012), soil organic matter dynamics as well as the distribution and density of soil microbes (Kong et al., 2009). Several efforts have been made towards the spatial prediction of particle size fraction (PSFs). Nemes et al., (1999) used four different interpolation techniques (Loglinear, Gompetz curve, non-parametric spline function and similarity indices) to study the spatial distribution patterns of PSFs. Scull et al., (2005b) compared the use of several statistical and geostatistical models to predict PSFs. Santra et al., (2008) also studied the spatial variation of PSFs using ordinary kriging. More recently, Adhikari et al (2013) employed regression tree approach to predict the PSFs for Denmark with the use of covariates such as DEM, land use, parent material, etc. Although all these studies clearly showed the significance of soil PSFs, none of these studies considered the compositional nature of PSFs.

One major challenge in operational DSM is the spatial prediction of soil particle size fractions as compositional data (Buchanan et al., 2012), such that the three component fractions have to sum to a constant, with distributions that are curtailed at the limits of 0 and 100. According De Gruijter et al. (1997) composition data must meet the following criteria:

a. Each of the components of the composition must be non-negative

$$Z * ij(x) \geq 0 \qquad [2.3]$$

where $Z * ij(x)$ is the estimate of a compositional regionalized variable, of

jth component at ith location.

b. At each location, the components must sum to a constant

$$\sum_{j=1}^{n} Z * ij(x) = \emptyset, \text{ and } \emptyset\text{=constant.} \qquad [2.4]$$

c. Estimates of the composition should be unbiased

$$Z^*{}_{ij}(x) = \sum_{i=1}^{n} \lambda_i Z_{ij}(x) \ \sum_{i=1}^{n} \lambda_i = 1; j = 1, \dots, k. \qquad [2.5]$$

Out of k components in the composition, $Z^*{}_{ij}(x)$ represents the estimate of a compositional regionalized variable, of the jth component at the ith location.

For the interpretation of regionalized compositions the sample space is a positive (S^d) and not a multidimensional space (R^d) (Aitchison, 1990). A d-part simplex is thus defined as:

$$S^d = \left\{x = [x_1, x_2, \dots\dots, x_d];\ x_i > 0, i = 1,2, \dots\dots\dots, d;\ \sum_{i=1}^{d} x_i = k \right\} \qquad [2.6]$$

where S^d represent row vectors of d-part compositions; k is a constant, which is the sum of vectorial compositions which could be 100 (if composition is a percentage) or 1. The transformation of this simplex S^d to the real space R^d, can be achieved using three (3) different approaches; additive log-ratio (Aitchison, 1990), centred log-ratio (Aitchison, 2003) and isometric log-ratio (Egozcue et al., 2003). According to Aitchison (1990) the additive log-ratio (ALR) can be expressed as:

$$y_{ij}(x) = ln \frac{z_{ij}(x)}{z_{ik}(x)} \quad k = d + 1 \quad i = 1, \dots., n \qquad [2.7]$$

where y_{ij} is the log ratio transformation of z_{ij}.

The inverse transformation of the above equation is:

$$z_{ij}(x) = \frac{\exp y_{ij}(x)}{\sum_{j=1}^{k} \exp y_k(x)} \qquad [2.8]$$

This compensates for the closure effect and subsequently through perturbation, the transformed data may fit a normal distribution, making the data suited to classical analysis such as MLR (Odeh et al., 2003).

One major criticism for the use of ALR has been the choice of an arbitrary component of the composition as a divisor. This, according to earlier critics, is problematic in the sense that, the distances between points in the transformed space are not the same for different divisors. However, it has been proven that linear statistical methods with compositional data as the dependent variable are invariant to the choice of divisor as the implicit linear transformations between different representations cancel out in any F ratio of quadratic or bilinear forms (Aitchison et al., 2000). Moreover, of the three log-ratio transformation methods, ALR has gained more usage in scientific research communities because of its ease of interpretation. It has been used in Soil Science studies for prediction of particle size fractions (PSFs) with predicted PSFs satisfying the criterion for compositional data analysis (Odeh et al., 2003, Buchanan et al., 2012; Li et al., 2012 Arrouays et al., 2011; Huang et al., 2014; Sun et al., 2014).

2.9 Estimation of total soil carbon stock

Since total carbon stock is one of the key functional properties derived in chapter 4 of this thesis it will be nice to review available approaches to soil carbon stock estimation. Generally, there are two major techniques used for estimating SOC stocks. These include the DSM and the measure-and-multiply (MM) (Mishra et al., 2010; Thompson and Kolka, 2005). The DSM approach estimates the spatial variability of SOC stocks in relation to variations in a set of environmental covariates (Mishra et al., 2010, Cambule et al., 2014; Were et al., 2015). Thereafter, predicted SOC stocks at the various grid cells are summed up to the total area (Gessler et al., 2000; Thompson et al., 2001). The use of dense spatial attributes accrued the DSM approach an advantage of relatively lower estimation error at each prediction location than the MM approach. Several studies have used DSM approach to estimate SOC stock (Florinsky et al., 2002; Ziadat, 2005; Ungaro et al., 2010; Adhikari et al., 2014; Cambule et al., 2014; Dorji et al., 2014; Were et al., 2015). However, many of these studies cover only small areas of about 100 hectares and only Adhikari et al. (2014) covers national scale application.

On the other hand, the MM approach begins with stratification of the entire study area after which point SOC stock estimates per stratum are averaged and multiplied by the area of each stratum (Thompson and Kolka, 2005; Guo et al., 2006; Tan et al., 2009). This approach has been applied to a numerous SOC stock studies ranging from regional to global scales (Amichev and Galbraith, 2004; Tan et al., 2004; Thompson and Kolka, 2005; Batjes, 2008). It has an advantage of simplicity and ease of use compared to the SLM approach. However, it is criticized by the possibility of high estimate error due to high within-stratum SOC variability (Thompson and Kolka, 2005; Mishra et al., 2010).

2.10 Deriving additional value from DSM products for National planning purposes

As an economic product, soil information has little or no value until it is interpreted and applied in such a way as to support decision making process (Grealish et al., 2015). In this context, primary DSM products can be used in quantitative land suitability assessment for crop production, irrigation needs and scheduling as well as land degradation assessment (Omuto et al., 2013). The outputs of such assessments are integral components of national agricultural and environmental planning; informing farmers and policy makers on where is best for the production of a particular crop or whether land could be allocated for alternative uses.

2.10.1 Land suitability assessment

Land suitability assessment involves the use of various soil parameters (such as soil texture, water retention capacity, exchangeable cation), along with other climate and terrain data, to identify suitable areas for various agricultural enterprises such as irrigation agriculture. Land suitability assessment is an integral component of developmental planning in most developing parts of the world, especially in the SSA where there is high demand for food and fibre. Basically, there are two major approaches in land suitability assessment. These include qualitative assessment which is based on expert judgment and quantitative assessment based on parametric method and process-oriented simulation models (Van Lanen, 1991; Bouma et al., 1993). In both approaches, the results are usually presented as maps where the class limits are based on rigid and exact data models (Burrough, 1992; McBratney and Odeh, 1997; Triantafilis et al., 2001) with the assumption that the structure and the parameters of the model are known with high certainty of occurrence (Zimmermann, 1992). However, according to Zadeh’s (1965) report on theory of fuzzy sets, almost all classes of objects encountered in the real physical world do not have precise criteria of membership. As discussed above in section 2.6.2.2, fuzzy set approach can be considered as an alternative method to crisp models in land evaluation (McBratney and Odeh, 1997). The main advantage of fuzzy sets is their capability to express gradual transitions from membership to non-membership and vice versa (Klir, 1995). However, the performance of fuzzy set methods depends mainly on membership function information (Zimmermann, 1992).

Whether a fuzzy or crisp approach is used, land suitability assessment involves complex interactions of biophysical, chemical and climatic processes with socioeconomic factors. These processes and factors are in most cases heterogeneous, interdependent and conflicting in nature. While the biophysical elements tend to be relatively stable, socio-economic factors are dynamic and dependent on the prevailing social, economic and political conditions of an area (Triantafilis et al., 2001; Keshavarzi et al., 2010). Therefore aggregating such heterogeneous criteria for decision making poses a major challenge for suitability assessment. To overcome these challenges, fuzzy decision making uses several aggregation operators on fuzzy sets for obtaining different types of decision functions.

2.10.2 Aggregation methods for fuzzy land suitability assessment

Decisions on suitability of land for a particular use, like most real life decision making, borders on making decisions under multiple attributes or evaluation criteria and multiple objective optimization. These are usually achieved through aggregating preferences, obtained from different decision makers on a given set of alternatives. Therefore, fuzzy decision making uses several aggregation operators on fuzzy sets to obtain different types of decision functions. When an array of evaluation criteria and the corresponding weight matrix for a particular decision are determined, information about the evaluation criteria is combined through an aggregation function to determine the overall suitability (Soasa and kaymak, 2002). The aggregation function may consist of a single aggregation operator or a combination of operators.

Zaheh (1965) introduced the first standard fuzzy operators; union (Max), intersection (Min), and complement which extensions of OR, AND, and NOT logical Boolean operations, respectively. However these standard aggregators do not express satisfactorily, the degree of compensation common to human aggregate criteria (Peneva, 2003). Therefore in fuzzy decision making other types of aggregations that are commonly used include conjunctive, disjunctive and compensatory aggregations (Soasa and Kaymuc, 2002). Conjunctive aggregation of criteria (t-norms) implies simultaneous satisfaction of all decision criteria, while the disjunctive aggregation (t-conorms) implies full compensation amongst the criteria. The compensatory aggregation (averaging operators such as OWA) is more suitable for dealing with conflicting criteria common with human aggregation behavior. However, it does not capture well enough the degree of compensation common to human aggregation ability in the presence of conflicting criteria.

Among the family of aggregation operators, fuzzy integrals are known to be one of the most robust aggregation functions that allow the fusion of information from several conflicting criteria (Torra and Narukawa, 2006). Fuzzy integral is based on the concept of fuzzy measure, which is a generalization of specific types of averaging aggregation operators (Grabisch et al., 2008). There are several fuzzy integrals: Choquet integral (Choquet, 1954), Sugeno integral (Sugeno, 1974), t-conorm integral (Murofushi and Sugeno, 1991) and twofold integral (Torra, 2003). Among these integrals, Choquet integral (CI) is one of the most commonly used for suitability analysis (Wang et al., 2006; Grabisch et al., 2008). It is non-linear, flexible based on either non-additive (Rowley et al., 2015) and/or additive measure. One important feature of CI is the capacity to recognize the vagueness of the decision environment and to account for the interactions among conflicting and correlated criteria (Yang, 2005). CI also considers the degree of satisfaction and/or dissatisfaction of alternatives for each criterion with the help of intuitionistic fuzzy values. Despite the important features of CI, it has been rarely used in soil suitability assessment (Odeh and Crawford, 2009; Chakan et al., 2012).

2.11 Application of DSM products for National scale planning

DSM products such as SOC and stock maps are useful components of national environmental monitoring programs (Minasny et al., 2013). Digital SOC and stock maps can be used in modelling spatio-temporal trends of soil processes in response to land use change. Such information will help national policy makers to develop plans on alternative land-management techniques. For instance, Milne et al. (2007) used SOC data in the Global Environment Facility Soil Organic Carbon modelling system to map future SOC stock changes in Brazilian Amazon (Cerri et al., 2007), the Indo-Gangetic plains (Bhattacharyya et al., 2007), and Jordan (Al-Adamat et al., 2007). Another application of DSM products for national planning is in the area of irrigation development planning. Irrigation development is a capital intensive project and as such, quantitative irrigation suitability maps which are derivatives of DSM products and landscape attributes can help guide national policy decisions-making on effective and economically viable irrigation establishment. Furthermore value-added DSM products such as soil fertility index, phosphorus retention capacity, pH buffering capacity are useful for national agricultural planning in terms of crop suitability. However, in most developing countries, particularly in the SSA, the lack of quantitative soil data has hampered application of soil information as described above for national planning.

2.12 Examples of operational DSM using legacy soil data

As previously stated, legacy soil data form the foundation or building blocks for most DSM operations (McBratney et al., 2003; Onuto et al 2013), especially in data-scarce countries. This is because of the need in such a situation to reduce the cost and difficulties in obtaining new samples for DSM. Thus legacy soil data can be used for model building and testing to produce soil information on previously unmapped areas (Hengl et al., 2015), for establishing areas of high uncertainty where new samples will be required to fill the gaps (Grimm and Brehens, 2010) and as baseline for studying change in soil properties over time (Karunaratne et al 2014). Another application of legacy soil data in DSM includes updating older soil survey information (Bui et al 2003; Kempen et al., 2015; Sun et al., 2015). Despite the availability of legacy soil data, only a few operational DSM studies at at the National and continental scales have utilized this useful source of data. Hong et al. (2009) mapped soil carbon storage and water capacity for Korea using legacy soil data from the Korean soil information system. Odgers et al. (2012) employed the weighted average approach to developing SOC map for the United States. Hengl et al. (2015) utilized the Africa soil database to map soil properties in the SSA. While these studies have succeeded in mapping basic soil properties at the different scales, none attempted adding value to the predicted soil attributes to support decision making and developmental planning especially at the national level.

2.13 Conclusion

There is a dire need for soil information to support increased production of food and fibres for a rapidly growing world population. This is particularly important in the SSA which is experiencing a continual increase in population that require more food, fibre and energy to be produced; against the background of the changing climate conditions and decline in land productivity in this region. However, soil information is very difficult to obtain, especially at more refined levels of detail. In addition, substantial financial investment is often required to obtain soil data because most soil properties exhibit high variability over short distances, so much that the skills and expertise necessary to accurately record, measure and map such changes are prohibitive in terms of time and labour. DSM can substantially help in providing soil information at the required format and scale in developing countries, particularly in SSA. However the practicability of DSM in these countries is limited by absence of dense spatial soil data.

There are several techniques such as data mining tools and hybrid models that are amendable to sparse legacy soil data. Currently, not much work has been done to apply these techniques to DSM in SSA especially at the national scale. Furthermore, most DSM operations are committed to producing digital maps or information on primary soil attributes. Such maps in themselves have very little or no value unless they are interpreted and applied to a particular question to support a decision-making process. One practical way of using primary soil attributes in planning and decision-making process, particularly at the national scale, is by incorporating them into multi-criteria suitability analysis for different land uses such as irrigation and plantation agriculture. Despite the established need for soil information especially in combating the major global issues, there has not been much effort to add additional value to DSM products to allow their effective utilization in decision making and planning at the national level.

References

Aalders, I., Hough, R. L., & Towers, W. 2011. Risk of erosion in peat soils–an investigation using Bayesian belief networks. *Soil Use and Management*, *27*(4), 538-549.

Abbasi, Y., Ghanbarian-Alavijeh, B., Liaghat, A. M., & Shorafa, M. 2011. Evaluation of pedotransfer functions for estimating soil water retention curve of saline and saline-alkali soils of Iran. *Pedosphere*, *21*(2), 230-237.

Adhikari, K., Kheir, R. B., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A.B., & Greve, M. H. 2013. High-resolution 3-D mapping of soil texture in Denmark. *Soil Science Society of America Journal*, *77*(3), 860-876.

Adhikari, K., Hartemink, A. E., Minasny, B., Kheir, R. B., Greve, M. B., & Greve, M. H. 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. *PloS ONE*, *9*(8), e105519.

Al Majou, H., Bruand, A., & Duval, O. 2008a. The use of in situ volumetric water content at field capacity to improve the prediction of soil water retention properties. *Canadian Journal of Soil Science*, *88*(4), 533-541.

Al Majou, H., Bruand, A., Duval, O., Le Bas, C., & Vautier, A. 2008b. Prediction of soil water retention properties after stratification by combining texture, bulk density and the type of horizon. *Soil Use and Management*, *24*(4), 383-391.

Amini, M., Afyuni, M., Fathianpour, N., Khademi, H., & Fluhler, H., 2005. Continuous soil pollution mapping using fuzzy logic and spatial interpolation. *Geoderma*, 124, 223-233.

Arrouays, D., Saby, N. P., Thioulouse, J., Jolivet, C., Boulonne, L., & Ratié, C. 2011. Large trends in French topsoil characteristics are revealed by spatially constrained multivariate analysis. *Geoderma*, *161*(3), 107-114.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J. Mendonça Santos, M.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., & Zhang, G. L. 2014. Chapter Three-GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. *Advances in Agronomy*, *125*, 93-134.

Aitchison, J. 2003. The statistical analysis of compositional data. In: Monographs on Statistics and Applied Probability. Blackburn Press, Caldwell, NJ.

Aitchison, J. 1990. Relative variation diagrams for describing patterns of compositional variability. *Mathematical Geology*, 22: 487–511.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. 2000. Logratio analysis and compositional distance. *Mathematical Geology*, *32*(3), 271-275.

Baker, L., & Ellison, D., 2008. Optimization of pedotransfer functions using an artificial neural network ensemble method. *Geoderma*. 144, 212–224.

Barthold, F. K., Stallard, R.F., & Elsenbeer, H., 2008. Soil nutrient-landscape relationships in a lowland tropical rainforest in Panama. *Forest Ecology and Management*, 255, 1135–1148. Doi:10.1016/j.foreco.2007.09.089

Bastet, G., Bruand, A., Voltz, M., Bornand, M. & Que´tin, P., 1997. Performance of available pedotransfer functions for predicting the water retention properties of French soils, Proceedings of International. Workshop on the Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media. University of California, Riverside, CA.

Batjes, N. H. 2008. Mapping soil carbon stocks of Central Africa using SOTER. *Geoderma*, *146*(1), 58-65.

Behrens, T. & Scholten, T. 2006. A comparison of data-mining techniques in predictive soil mapping. In: Lagacherie, P., A.M., Voltz, M. (Eds.), Digital Soil Mapping—an Introductory Perspective. Developments in Soil Science. Elsevier, volume 31 pp. 353–364.

Bell, M.A., & van Keulen, H. 1996. Effect of soil disturbance on pedotransfer function development for field capacity. *Soil Technology*. 8: 321-329.

Bell, J. C., Cunningham, R. L., & Havens, M. W. 1994. Soil drainage class probability mapping using a soil-landscape model. *Soil Science Society of America Journal*, *58*(2), 464-470.

Bishop, T.F.A., A.B. McBratney and G.M Laslett. 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, 91, 27–45, http://dx.doi.org/10.1016/S0016-7061(99)00003-8.

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. *Advances in Soil Science* 9: 177–213.

Bouma, J., Wagenet, R.J., Hoosbeek, M.R., & Hutson, J.L. 1993. Using expert systems and simulation modelling for land evaluation at farm level: a case study from New York State. *Soil Use and Management*, 9, 131-139.

Botula, Y. D., Cornelis, W. M., Baert, G., & Van Ranst, E. 2012. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (DR Congo). *Agricultural Water Management*, *111*, 1-10.

Botula, Y. D., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J., & Cornelis, W. M. 2015. Hierarchical Pedotransfer Functions to Predict Bulk Density of Highly Weathered Soils in Central Africa. *Soil Science Society of America Journal*, *79*(2), 476-486.

Braimoh, A. K., Vlek, P. L. G., & Stein, A. 2004. Land evaluation for maize based on fuzzy set and interpolation. *Environmental Management*, 33(2), 226–238.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). Classification and Regression trees- Wadsworth & Brooks Wadsworth statistics/probability series.

Breiman, L. (2001a). Random forests. *Machine Learning* 45(1):5–32.

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

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

Buchanan, S., Triantafilis, J., Odeh, I., & Subansinghe, R. 2012. Digital soil mapping of compositional particle-size fractions using proximal and remotely sensed ancillary data. *Geophysics*, 77(4), 201-211.

Bui, E. N., Henderson, B. L., & Viergever, K. 2006. Knowledge discovery from models of soil properties developed through data mining. *Ecological Modelling*, *191*(3), 431-446.

Burrough, P.A., & McDonell, R.A., 1998. Fuzzy classification methods for determining land suitability from soil profile observation and topography. *Journal of Soil Sciences*, 43, 193-210.

Calhoun, F. G., Smeck, N. E., Slater, B. L., Bigham, J. M., & Hall, G. F. 2001. Predicting bulk density of Ohio soils from morphology, genetic principles, and laboratory characterization data. *Soil Science Society of America Journal*, *65*(3), 811-819.

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.

Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J., & Smaling, E. M. A. 2015. Rescue and renewal of legacy soil resource inventories: A case study of the Limpopo National Park, Mozambique. *Catena*, *125*, 169-182.

Carré, F., McBratney, A. B., Mayr, T., & Montanarella, L. 2007b. Digital soil assessments: Beyond DSM. *Geoderma*, *142*(1), 69-79.

Chakan, A.A., Sarmadian, F., Heydari, A., Omid, M., Mohammadi, J., & Odeh, I. 2012. Performance assessment of fuzzy integrals as aggregation operators for soil suitability analysis (Case Study: Silakhor Plain of Lorestan Province). *Journal of Range and Watershed Management,* 65 (3), 301-314.

Collard, F., Saby, N.P.A., de Forges, A. H., Lehmann, S., Paroissien, J.B., & Arrouays, D. 2014. Spatial prediction of soil organic carbon at different depths using digital soil mapping. In: Arrouays, D., McKenzie, N., Hempel, J., de Fprges, A.R., McBratney, A.B. (EDs.). GlobalSoilMap: Basis of the global spatial soil information system, p 181-184. CRC press.

Cook, S.E., Jarvis, A. & Gonzalez, J.P. 2008. A new global demand for digital soil information. In: Hartemink, A.E., A.B. McBratney, M.L. Mendonca-Santos, M.L. (eds.). Digital Soil Mapping with Limited Data. Springer. 445p.

Craig, E., & Huettmann, F. 2008. Using 'blackbox' algorithms such as Treenet and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data: An overview, an example using Golden Eagle satellite data and an outlook for a promising future. In: H. F. Wang (Ed.). Intelligent data analysis: Developing new methodologies through pattern discovery and recovery. IGI Global, Hershey, PA, USA. P 65-84.

Cresswell, H. P., Coquet, Y., Bruand, A., & McKenzie, N. J. (2006). The transferability of Australian pedotransfer functions for predicting water retention characteristics of French soils. *Soil Use and Management, 22*(1), 62-70.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. 2007. Random forests for classification in ecology. *Ecology, 88*(11), 2783-2792.

Davidson, D.A., Theocharopoulos, S.P., Bloksma, R.J., 1994. A land evaluation project in Greece using GIS and based on Boolean and fuzzy set methologies. *International Journal of Geographical Information Systems*, 8(4), 369-384.

De Gruijter, J. J., Walvoort, D. J. J., & Van Gams, P. F. M. 1997. Continuous soil maps—a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma*, *77*(2), 169-195.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., & Muys, B. (2005). Predictive quality of pedotransfer functions for estimating bulk density of forest soils. *Soil Science. Society of America Journal,* 69,500–510. doi:10.2136/sssaj2005.0500

Dlamini, W. M. 2010. A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. *Environmental Modelling & Software*, *25*(2), 199-208.

Dobos, E., Carré, F., Hengl, T., Reuter, H. I., & Tóth, G. 2006. Digital soil mapping as a support to production of functional maps. *Luxemburg: Office for Official Publications of the European Communities, EUR, 22123*. Luxemburg. 68 p.

Dorji, T., Odeh, I. O., Field, D. J., & Baillie, I. C. 2014. Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems, Eastern Himalayas. *Forest Ecology and Management*, *318*, 91-102.

Efron, B., & Tibshirani, R. J. 1994. *An introduction to the bootstrap*. CRC press, London. 456p.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barcelo-Vidal, C. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology,* 35(3), 279-300.

Elith, J., Leathwick, J. R., & Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802-813.

Finke, P. A. 2012. On digital soil assessment with models and the Pedometrics agenda. *Geoderma, 171-172,* 3-15.

Florinsky, I. V., Eilers, R. G., Manning, G. R., & Fuller, L. G. 2002. Prediction of soil properties by digital terrain modelling. *Environmental Modelling & Software*, *17*(3), 295-311.

Friedman, J.H., & Meulman, J.J. 2003. Multiple additive regression trees with application in view of boosting. *The Annals of Statistics*, 28, 337-407.

Genske, D. D., Heinrish, K., 2009. A knowledge-based fuzzy expert system to analyse degraded terrain. *Expert Systems with Applications*, 36, 2459-2472.

Gessler, P.E., Chadwick, O.A., Chamran, F., Althouse, L., Holmes, K., 2000. Modeling soillandscape and ecosystem properties using terrain attributes. *Soil Science Society of America Journal* 64 (6), 2046–2056.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. 2006. Random forests for land cover classification. *Pattern Recognition Letters*, *27*(4), 294-300.

Goovaerts, P. 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, p. 483.

Grealish, G. J., Fitzpatrick, R. W., & Hutson, J. L., 2015. Soil survey data rescued by means of user friendly soil identification keys and toposequence models to deliver soil information for improved land management. *Geographical Research Journal*, *6*, 81-91.

Grimm, R., & Behrens, T. 2010. Uncertainty analysis of sample locations within digital soil mapping approaches. *Geoderma*, 155(3), 154-163.

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H., (2008). Soil organic carbon and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis. *Geoderma*, *146*(1), 102-113.

Grunwald, S., Thompson, J. A., & Boettinger, J. L. 2011. Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Science Society of America Journal*, *75*(4), 1201-1213.

Grunwald, S., Vasques, G. M., & Rivero, R. G. 2015. Chapter One-Fusion of Soil and Remote Sensing Data to Model Soil Properties. *Advances in Agronomy*, *131*, 1-109.

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*, 46-54.

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

Hartemink, A.E., Kienast-Brown, S. (Eds.) 2010. Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer. 439p.

Hartemink, A.E., Krasilnikov, P., Bockheim, J., 2013. Soil maps of the world. *Geoderma*, 207, 256-267.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. The Elements of Statistical Learning. 2nd edition. Springer, Berlin 739p.

Henderson, B. L., Bui, E. N., Moran, C. J., & Simon, D. A. P. 2005. Australia-wide predictions of soil properties using decision trees. *Geoderma*, *124*(3), 383-398.

Hansen, M. K., Brown, D. J., Dennison, P. E., Graves, S. A., & Bricklemyer, R. S. 2009. Inductively mapping expert-derived soil-landscape units within dambo wetland catenae using multispectral and topographic data. *Geoderma*, *150*(1), 72-84.

Hempel, J.W., Hammer, A.C., Moore, R.D., Bell, J.C., Thompson, J.A., & Golden, M.L. 2008. Challenges to digital soil mapping. In: Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (eds.). Digital Soil Mapping with Limited Data. Springer. 445p.

Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. 2007. About regression-kriging: From equations to case studies. Computers and Geosciences, 33(10), 1301-1315.

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L. & Tondoh, J. E. 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PloS one*, *10*(6), e0125814.

Herr, H. D., & Krzysztofowicz, R. 2015. Ensemble Bayesian forecasting system Part I: Theory and algorithms. *Journal of Hydrology*, *524*, 789-802.

Heuscher, S. A., Brandt, C. C., & Jardine, P. M. 2005. Using soil physical and chemical properties to estimate bulk density. *Soil Science Society of America Journal*, *69*(1), 51-56.

Heuvelink, G. B. M., Hoogland, T., & Walvoort, D. J. J. 2010. *A disposition of interpolation techniques* (No. 190, p. 86). Wettelijke Onderzoekstaken Natuur & Milieu.

Hong, S. Y., Minasny, B., Zhang, Y. S., Kim, Y. H., & Song, K. C. 2009. Digital soil mapping of soil carbon storage and available water capacity in Korea. In Proceedings of the 9th International Conference of The East and Southeast Asia Federation of Soil Sciences Societies, Seoul (pp. 406-408).

Hough, R. L., Towers, W., & Aalders, I. 2010. The risk of peat erosion from climate change: land management combinations—an assessment with bayesian belief networks. *Human and Ecological Risk Assessment*, *16*(5), 962-976.

Hu, Z., Chan, C.W., & Huang, G.H., 2003. A fuzzy expert system for site characterization. *Expert Systems with Applications*, 24, 123-131.

Huang, J., Subasinghe, R., & Triantafilis, J. 2014. Mapping Particle-Size Fractions as a Composition Using Additive Log-Ratio Transformation and Ancillary Data. *Soil Science Society of America Journal*, *78*(6), 1967-1976.

Jain, M. K., Kothyari, U. C., & Raju, K. G. 2005. GIS based distributed model for soil erosion and rate of sediment outflow from catchments. *Journal of Hydraulic Engineering*, *131*(9), 755-769.

Jalabert, S. S. M., Martin, M. P., Renaud, J. P., Boulonne, L., Jolivet, C., Montanarella, L., & Arrouays, D. 2010. Estimating forest soil bulk density using boosted regression modelling. *Soil Use and Management*, *26*(4), 516-528.

Jafarzadeh, A. A., Pal, M., Servati, M., FazeliFard, M. H., & Ghorbani, M. A. 2015. analysis of support vector machine and artificial neural network models for soil cation exchange capacity prediction. *International Journal of Environmental Science and Technology*, 1-10.

Joss, B. N., Hall, R. J., Sidders, D. M., & Keddy, T. J. 2008. Fuzzy-logic modeling of land suitability for hybrid poplar across the Prairie Provinces of Canada. *Environmental Monitoring and Assessment*, *141*(1-3), 79-96.

Karunaratne, S. B., Bishop, T. F. A., Odeh, I. O. A., Baldock, J. A. and Marchant, B. P. 2014. Estimating change in soil organic carbon using legacy data as the baseline: issues, approaches and lessons to learn. *Soil Research* 52(4), 349-365.

Kaur, R., Kumar, S., & Gurung, H. P. 2002. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs. *Soil Research*, *40*(5), 847-858.

Kempen, B., Brus, D. J., & de Vries, F. 2015. Operationalizing digital soil mapping for nationwide updating of the 1: 50,000 soil map of the Netherlands. *Geoderma*, 241, 313-329.

Keshavarzi, A., Sarmadian, F., Heidari, A., & Omid, M. 2010. Land suitability evaluation using fuzzy continuous classification (a case study: Ziaran region). *Modern Applied Science*, *4*(7), p72.

Kettler, T.A., Doran, J.W., & Gilbert, T.L. 2001. Simplified method for soil particle-size determnination to accompany soil-quality analyses. *Soil Science Society of America Journal*, 65, 849-852.

Kidd, D. B., Webb, M. A., Grose, C. J., Moreton, R. M., Malone, B. P., McBratney, A. B., Minasny, B., Viscarra-Rossel, R.A., Cotching, W.E., Sparrow, L.A., & Smith, R. 2012. Digital soil assessment: Guiding irrigation expansion in Tasmania, Australia. In: Minasny, B., Malone, B.P., McBratney, A. (Eds.). Digital Soil Assessments and Beyond. Taylor & Francis Group, London, 482p

Kim, J., Mohanty, B. P., & Shin, Y. 2015. Effective Soil Moisture Estimate and its Uncertainty using Multi-Model Simulation based on Bayesian Model Averaging. *Journal of Geophysical Research: Atmospheres*.

Klir, G., & Yuan, B. 1995. Fuzzy sets and fuzzy logic: theory and applications. Volume 4. Prentice Hall, New Jersey.

Kong, X., Dao, T.H., Qin, J., Li, C., & Zhang, F. 2009. Effects of soil texture and land use interactions on organic carbon in soils in North China cities's urban fringe. *Geoderma*, 105, 49-80.

Kuhnert, P. M. 2011. Four case studies in using expert opinion to inform priors. *Environmetrics*, *22*(5), 662-674.

Kuhnert, P. M., Martin, T. G., & Griffiths, S. P. 2010. A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecology Letters*, *13*(7), 900-914.

Lagacherie, P. 1992. *Formalisation des lois de distribution des sols pour automatiser la cartographie pédologique à partir d'un secteur pris comme référence* (Doctoral dissertation).

Lagacherie, P., & Holmes, S. 1997. Addressing geographical data errors in a classification tree for soil unit prediction. *International Journal of Geographical Information Science*, 11(2), 183-198.

Lagacherie, P. (2005). An algorithm for fuzzy pattern matching to allocate soil individuals to pre-existing soil classes. *Geoderma*, *128*(3), 274-288.

Lake, H. R., Akbarzadeh, A., & Mehrjardi, R. T. 2009. Development of pedotransfer functions (PTFs) to predict soil physico-chemical and hydrological characteristics in southern coastal zones of the Caspian Sea. *Journal of Ecology and the Natural Environment*, *1*(7), 160-172.

Lal, R., 2008. Soils and sustainable agriculture. A review. *Agronomy for Sustainable Development*, 28(1), 57-64.

Lamorski, K., Pachepsky, Y., Sławiński, C., & 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.

Lawrence, R., Bunn, A., Powell, S., Zambon, M. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90 (3), 331–336.

Lawrence, R.L., Wood, S.D., Sheley, R.L. 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (Random Forest). *Remote Sensing of Environment*, 100, 356-362.

Lemercier, B., Arrouays, D., Follan, S., Saby, N.P.A., Schvartz, C., Walter, C. 2008. Broad-scale soil monitoring through a nationwide soil-testing database. In: Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (eds.). Digital Soil Mapping with Limited Data. Springer. 445p.

Liaw, A., & Wiener, M., 2002. Classification and regression by random forest. *R News*, 2 (3), 18–22.

Li, J., & Heap, A. 2008. A Review of spatial interpolation methods for environmental Scientists No. Record 2008/23. Geoscience Australia, Canberra.

Li, J., & Heap, A. 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics,* 6, 228–241.

Li, J., Heap, A. D., Potter, A., & Daniell, J. J. 2011. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, *6*(12), 1647-1659.

Li, C. X., Luo, Y., Bao, A. M., Zhang, Y., Yang, C. J., & Cui, L. L. 2012. Study on spatial interpolation of compositional data based on log-ratio transformation. *Scientia Agricultura Sinica*, *45*(4), 646-655.

Ließ, M., Glaser, B., & Huwe, B. 2012. Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. *Geoderma*, *170*, 70-79.

Lin, H., J. Bouma, Y. Pachepsky, A. Western, J. Thompson, R. van Genuchten, H.-J. Vogel, and A. Lilly 2006. Hydropedology: Synergistic integration of pedology and hydrology, *Water Resources Research. 42(5)*, W05301, doi: 10.1029/2005WR004085.

Lloyd C.D., 2005. Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain. *Journal of Hydrology* 308:128–150

Loh, W. Y. 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14-23.

Lorenzetti, R., Barbetti, R., Fantappiè, M., L'Abate, G., & Costantini, E. A. 2015. Comparing data mining and deterministic pedology to assess the frequency of WRB reference soil groups in the legend of small scale maps. *Geoderma*, *237*, 237-245.

MacMillan, R. A. 2008**.** Experiences with Applied DSM: Protocol, Availability, Quality and Capacity Building. In: Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (Eds.). Digital Soil Mapping with Limited Data. Springer. 445p.

Malins, D., & Metternicht, G. 2006. Assessing the spatial extent of dryland salinity through fuzzy modelling. *Ecological Modellin*g, 193 (3), 387-411.

Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, *154*(1), 138-152.

Manrique, L. A., & Jones, C. A. 1991. Bulk density of soils in relation to soil physical and chemical properties. *Soil Science Society of America Journal*, *55* (2), 476-481.

Manyame, C., Morgan, C.L., Heilman, J.L., Fatondji, D., Gerard, B., Payne, W.A., 2007. Modeling hydraulic properties of sandy soils of Niger using pedotransfer functions. *Geoderma*, 141, 407–415.

Martin, M.P., Seen, D.l., Boulonne, L., Jolivet, C., Nair, K.M., Bourgeon, G., Arrouays, D., 2009. Optimizing pedotransfer functions for estimating soil bulk density using boosted regression trees. *Soil Science Society of America Journal*, 73, 485–493.

Martin, M. P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., & Arrouays, D. (2011). Spatial distribution of soil organic carbon stocks in France. *Biogeosciences*, *8*(5), 1053-1065.

Martin, M. P., Orton, T. G., Lacarce, E., Meersmans, J., Saby, N. P. A., Paroissien, J. B., Jolivet, C., Boulonne., L. & Arrouays, D. 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. *Geoderma*, *223*, 97-107.

Mayr, T. R., Palmer, R. C., & Cooke, H. J. 2008. Digital soil mapping using legacy data in the Eden valley, UK. In *Digital Soil Mapping with Limited Data* (pp. 291-301). Springer Netherlands.

McBratney, A. B., & de Gruijter, J.J. 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades. Journal of Soil Science, 43(1), 159-175.

McBratney, A. B., & Odeh, I. O. 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, *77* (2), 85-113.

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

McBratney, A.B., Mendonc¸a Santos, M.L., & Minasny, B. 2003. On digital soil mapping. *Geoderma*, 117, 3–52.

McBratney, A. B., Minasny, B., & Rossel, R. V. 2006. Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. *Geoderma*, *136*(1), 272-278.

McBratney, A. B., Minasny, B., & Tranter, G. 2011. Necessary meta-data for pedotransfer functions. *Geoderma*, *160*(3), 627-629.

McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. *Geoderma*, 213, 203-213.

McCann, R. K., Marcot, B. G., & Ellis, R. 2006. Bayesian belief networks: applications in ecology and natural resource management. *Canadian Journal of Forest Research*, *36*(12), 3053-3062.

McCloskey, J. T., Lilieholm, R. J., & Cronan, C. 2011. Using Bayesian belief networks to identify potential compatibilities and conflicts between development and landscape conservation. *Landscape and Urban Planning*, *101*(2), 190-203.

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

Medina, H., Tarawally, M., del Valle, A., & Ruiz, M. E. 2002. Estimating soil water retention curve in rhodic ferralsols from basic soil data. *Geoderma*, *108*(3), 277-285.

Merdun, H., Çınar, Ö., Meral, R., & Apan, M. 2006. Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. *Soil and Tillage Research*, *90*(1), 108-116.

Merdun, H. 2010. Alternative methods in the development of pedotransfer functions for soil hydraulic characteristics. *Eurasian Soil Science*, 43(1): 62-71

Miller, B.A., Koszinski, S., Wehrhan, M., & Sommer, M., .2015. Impact of multi-scale predictor selection for modeling soil properties. *Geoderma*, 239–240, 97–106

Minasny, B., McBratney, A. B., & Bristow, K. L. 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma*, *93*(3), 225-253.

Minasny, B. & McBratney, A. B. 2002. The Neuro-m method for fitting neural network parametric pedotransfer functions. *Soil Science. Society of America Journal*, 66, 352–361.

Minasny, B., Hopmans, J.W., Harter, T., Eching, S.O., Tuli, A., & Denton, M.A. 2004. Neural networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data. *Soil Science Society of America Journal*, 68, 417–429.

Minasny B., McBratney, A.B., Mendoca-Santos, M.L., Odeh, I.O.A., & Guyon, B. 2006 Prediction and digital mapping of soil carbon storage in the lower Namoi valley. *Australian Journal of Soil Research*, 44**,** 233-244.

Minasny,B., McBratney, A.B., & Lark, R.M. 2008. Digital Soil Mapping Technologies for

Countries with Sparse Data Infrastructures. In: Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (eds.). Digital Soil Mapping with Limited Data. Springer. 445p.

Minasny, B., & Hartemink, A. E. 2011. Predicting soil properties in the tropics. *Earth-Science Reviews*, *106*(1), 52-62.

Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. 2013. Digital mapping of soil carbon. *Advances in Agronomy*, *118*, 1-47.

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

Mishra, U., Lal, R., Liu, D., Van Meirvenne, M., 2010. Predicting the spatial variation of the soil organic carbon pool at a regional scale. *Soil Science Society of America Journal*, 74 (3), 906–914.

Moran, J.M., & Bui, E.N. 2002. Spatial data mining for enhanced soil map modelling. *International Journal of Geographic Information Science*, 16 (6), 533–549.

Mora-Vallejo, A., Claessens, L., Stoorvogel, J. & Heuvelink, G.B.M. 2008. Small scale digital soil mapping in Southeastern Kenya. *Catena*, **7**6, 44–53.

Nemes, A., Wösten, J. H. M., Lilly, A., & Voshaar, J. O. 1999. Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. *Geoderma*, *90*(3), 187-202.

Nemes, A., & Rawls, W. J. 2006. Evaluation of different representations of the particle-size distribution to predict soil water retention. *Geoderma*, 132, 47–58.

Nemes, A., Rawls, W. J. & Pachepsky, Y. A. (2005). The influence of organic matter on the estimation of saturated hydraulic conductivity. *Soil Science Society of America Journal*, 69, 1330–1337.

Nemes, A., Rawls, 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.

Nemes, A., Quebedeaux, B., & Timlin, D.J. 2010. Ensemble approach to provide uncertainty estimates of soil bulk density. *Soil Science Society of America Journal*, 74, 1938–1945. doi:10.2136/sssaj2009.0370

Northcliff, S,. 2009. The soil: nature, sustainable use, management, and protection. An overview. *Gaia-Ecological Perspectives for Science and Society, 18(1), 58-68*

Odeh, I.O.A., McBratney, A.B., & Chittleborough, D.J., 1992. Fuzzy c-means and kriging for mapping soil as a continuous system. *Soil Science Society of America Journal,* 56, 1848-*1854.*

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

Odeh, I. O., McBratney, A. B., & Chittleborough, D. J. 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, 67(3), 215-226.

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

Odeh, I. O., Todd, A. J., & Triantafilis, J. 2003. Spatial prediction of soil particle-size fractions as compositional data. *Soil Science*, *168*(7), 501-515.

Odeh, I. & Crawford, M. 2009. On fuzzy measures, choquet integral and classical aggregation operators for multi-dimensional decision-making in soil suitability assessment. *Biennial Meeting of Commission 1.5 Pedometrics Division 1 of the International Union of Soil Science (IUSS) held* August 26-28. China Agricultural University Beijing, China.

Odeh, I.O.A., Leenaars, J.G.B., Hartemink, A., & Amapu, I. 2012. The challenges of collating legacy data for digital mapping of Nigerian soils. In: Minasny, B., Malone, .P., McBratney, A. (eds). Digital Soil Assessments and Beyond. Taylor & Francis Group, London, 482p.

Omuto, C., Nachtergaele, F., & Rojas, R. V. 2013. *State of the Art Report on Global and Regional Soil Information: Where are we? Where to go?* Food and Agriculture Organization of the United Nations. Rome. 81p.

Padarian, J., Minasny, B., McBratney, A. B., & Dalgliesh, N. 2014. Predicting and mapping the soil available water capacity of Australian wheatbelt. *Geoderma Regional*, *2*, 110-118.

Peneva, V. & Popchev, I. 2003. Properties of the aggregation operators related with fuzzy relations. *Fuzzy Sets and Systems* 139, 615-633.

Peters, J., Verhoest, N. E. C., Samson, R., Boeckx, P., & De Baets, B. 2008. Wetland vegetation distribution modelling for the identification of constraining environmental variables. *Landscape ecology*, *23*(9), 1049-1065.

Powers, J. S., Corre, M. D., Twine, T. E., & Veldkamp, E. 2011. Geographic bias of field observations of soil carbon stocks with tropical land-use changes precludes spatial extrapolation. *Proceedings of the National Academy of Sciences*, *108*(15), 6318-6322.

Prasad, A. M., Iverson, L.R., &Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9: 181-199.

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

Qi, F., Zhu, A.X., Pei, T., Qin, C., & Burt, J.E. 2008. Knowledge discovery from area-class resource maps: capturing prototype effects. *Cartography and Geographic Information Science*, 35(4), 223-237.

Rab, M.A., Chandra, S., Fisher, P.D., Robinson, N.J., Kitching, M., Aumann, C.D., & Imhoff, M. 2011 Modelling and prediction of soil water contents at field capacity and permanent wilting of dryland cropping soils. *Soil Research,* 49, 389–407. doi:10.1071/SR10160

Rawls, W.J., Nemes, A. & Pachepsky, Y.A. 2004. Effect of soil organic carbon on soil hydraulic properties. In: Development of pedotransfer functions in soil hydrology (eds Y. Pachepsky & W.J. Rawls), pp. 95–114. Elsevier, Amsterdam, The Netherlands.

Romano, N., & Chirico, G. B. 2004. The role of terrain analysis in using and developing pedotransfer functions. *Developments in Soil Science*, *30*, 273-294.

Rossiter, D. 2008. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructure. In: Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (eds.). Digital Soil Mapping with Limited Data. Springer. 445p.

Safari, Y., Boroujeni, I.E., Kamali, A., Salehi, M.H., & Bodaghabadi, M.B. 2013. Mapping of the soil texture using geostatistical method (a case study of the Shahrekord plain, central Iran). Arabian Journal of Geosciences, 6(9), 3331-3339.

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.D., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., & Zhang, G.L., 2009. Digital soil map of the world. *Science*, 325, 680-681.

Santra, P., Chopra, U. K., & 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.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, 1651-1686.

Schmidt, K., Behrens, T., Daumann, J., Ramirez-Lopez, L., Werban, U., Dietrich, P., & Scholten, T. 2014. A comparison of calibration sampling schemes at the field scale. *Geoderma*, *232*, 243-256.

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

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

Scull, P., Okin, G., Chadwick, O. A., & Franklin, J. 2005b. A comparison of methods to predict soil surface texture in an alluvial basin. *The Professional Geographer*, *57*(3), 423-437.

Sequeira, C. H., Wills, S. A., Seybold, C. A., & West, L. T. 2014. Predicting soil bulk density for incomplete databases. *Geoderma*, *213*, 64-73.

Sheng J D, Ma L C, Jiang P A, Li B G, Huang F., & Wu H Q. 2010. Digital soil mapping to enable classification of the salt-affected soils in desert agro-ecological zones. *Agricultural Water Management*, 97, 1961-1970.

Shi, X., Long, R. Dekett, R., & Philippe, J. 2009. Integrating Different Types of Knowledge for Digital Soil Mapping. *Soil Science Society of America Journal*, 73, 1682-1692.

Sicat, R. S., Carranza, E. J. M., & Nidumolu, U. B. 2005. Fuzzy modeling of farmers' knowledge for land suitability classification. *Agricultural Systems*, *83*(1), 49-75.

Smith, C. S., Howes, A. L., Price, B., & McAlpine, C. A. 2007. Using a Bayesian belief network to predict suitable habitat of an endangered mammal–The Julia Creek dunnart (Sminthopsis douglasi). *Biological Conservation*, *139*(3), 333-347.

Sousa, J.M.C. & Kaymak, U. 2002. Fuzzy Decision Making, Modeling and Control, World Scientific Publishing Company, 335p.

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.

Sun, X. L., Wu, Y. J., Wang, H. L., Zhao, Y. G., & Zhang, G. L. 2014. Mapping Soil Particle Size Fractions Using Compositional Kriging, Cokriging and Additive Log-ratio \ Cokriging in Two Case Studies. *Mathematical Geosciences*, *46*(4), 429-443.

Sun, X. L., Wu, Y. J., Lou, Y. L., Wang, H. L., Zhang, C., Zhao, Y. G., & Zhang, G. L. 2015. Updating digital soil maps with new data: a case study of soil organic matter in Jiangsu, China. *European Journal of Soil Science*, 66(6), 1012-1022.

Taalab, K., Corstanje, R., Mayr, T. M., Whelan, M. J., & Creamer, R. E. 2015a. The application of expert knowledge in Bayesian networks to predict soil bulk density at the landscape scale. *European Journal of Soil Science*.

Taalab, K., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M. J., Hannam, J. A., & Creamer, R. 2015b. On the application of Bayesian Networks in Digital Soil Mapping. *Geoderma*, *259*, 134-148.

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*, 67-77.

Thompson, J.A., Kolka, R.K., 2005. Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. *Soil Science Society of America Journal*, 69 (4), 1086–1093.

Thompson, J.A., Bell, J.C., Butler, C.A., 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*, 100 (1–2), 67–89.

Torbert, H.A., Krueger, E., & Kurtener, D. 2008. Soil quality assessment using fuzzy modelling. *International Agrophysics*, 22, 365-370.

Tranter, G., Minasny, B., McBratney, A.B., Murphy, B., McKenzie, N.J., Grundy, M. & Brough, D. 2007. Building and testing conceptual and empirical models for predicting soil bulk density. *Soil Use and Management*, 23, 437–443.

Tranter, G., Minasny, B., & McBratney, A. B. 2010. Estimating Pedotransfer Function Prediction Limits Using Fuzzy-Means with Extragrades. *Soil Science Society of America Journal*, *74*(6), 1967-1975.

Triantafilis, J., Ward, W. T., & McBratney, A. B. 2001. Land suitability assessment in the Namoi Valley of Australia, using a continuous model. *Soil Research*, *39*(2), 273-289.

Tveito, O.E. 2010. An assessment of circulation type classifications for precipitation distribution in Norway. *Physics and Chemistry of the Earth*, Parts A/B/C , 35(9), 395-402.

Twarakavi, N. K. C., Simunek, 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.

Ungaro, F., Calzolari, C. & Busoni, E. 2005. Development of pedotransfer functions using a group method of data handling for the soil of the Pianura Padanao-Veneta region of North Italy: Water retention properties. *Geoderma*, 124**,** 293–317.

Ungaro, F., Staffilani, F., Tarocco, P., 2010. Assessing and mapping topsoil organic carbon stock at regional scale: a scorpan kriging approach conditional on soil map delineations and land use. *Land Degradation & Development*, 21 (6), 565–581.

Vasques, G.M., Grunwald, S., Comerford, N.B., Sickman, J.O. 2010. Regional modelling of soil carbon at multiple depths within a subtropical watershed. Geoderma 156, 326–336.

Venables, W. N., & Ripley, B. D. 1994. *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M.G., & 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. doi:10.2136/vzj2010.0045

Verikas, A., Gelzinis, A., & Bacauskiene, M. 2011. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44, 330-349.

Viscarra, R., Rossel, A., McKenzie, N.J., & Grundy, M.J. 2010. Using Proximal Soil Sensors for Digital Soil Mapping. In: Boettinger, J.L., Howell, D.W., Moore, A.C.,

Vitharana, U.W.A., Saey, T., Cockx, L., Simpson, D., Vermeersch, H., & Van Meirvenne, M. 2008. Upgrading a 1/20,000 soil map with an apparent electrical conductivity survey.*Geoderma*, 148, 107–112.

Van Ranst, E., Verdoodt, A., & Bert, G., 2010. Soil mapping in Africa at the crossroads: work to make up forlost ground. *Bulletin des Séances d'Académie Royal des Sciences of d'Outre-Mer*, 56,147-163.

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators*, *52*, 394-403.

Yang, R., Wang, Z., Heng, P. A., & Leung, K. S. 2005. Fuzzy numbers and fuzzification of the Choquet integral. *Fuzzy Sets and Systems*, *153*(1), 95-113.

Zadeh, L.A., 1965. Fuzzy sets. *Information Control*, 8, 338-353.

Zhu, A. X., Band, L. E., Dutton, B., & Nimlos, T. J. 1996. Automated soil inference under fuzzy logic. *Ecological Modelling*, *90*(2), 123-145.

Zhu, A.X., Hudson, B., Burt, J., Lubish, K., & Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, 65(5), 1463-1472.

Ziadat, F.M., 2005. Analyzing digital terrain attributes to predict soil attributes for a relatively large area. *Soil Science Society of America Journal*, 69 (5), 1590–1599.

Zimmermann, H.J. 1992. Fuzzy Set Theory and its Applications, 2nd Edition, Kluwer Academic press, 399p.

Chapter 3.

Digital Mapping of Soil Particle Size Fractions at a National Scale

Published: **Akpa, S. I.C**, Odeh, I. O. A, Bishop, T. F.A, & Hartemink, A. E. (2014). Digital Mapping of Soil Particle-Size Fractions for Nigeria. *Soil Science Society of America Journal*, *78*(6), 1953-1966.

Abstract

There is a growing need for spatially continuous and quantitative soil information for environmental modeling and management especially at the national scale. This study was aimed at predicting soil particle-size fractions (PSF) for Nigeria using random forest model (RFM). Equal-area quadratic splines were fitted to Nigerian legacy soil profile data to estimate PSFs at six standard soil depths (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) using the GlobalSoilMap project specification. We applied an additive log-ratio (ALR) transformation of the PSFs. There was a better prediction performance (based on 33% model validation) in the upper depth intervals than the lower depth intervals (e.g., R^2 of 0.53; RMSE of 13.59 g $kg^{-1}$for clay at 0–5 cm and R^2 of 0.16; RMSE of 15.60 g kg^{-1} at 100–200 cm). Overall, the PSFs show marked variations across the entire Nigeria with a higher sand content compared with silt and clay contents and increasing clay content with soil depth. The variation in soil texture (ST) shows a progressive transition from a coarse texture (sand) along the fringes of northern Nigeria (e.g., upper part of Maiduguri and Sokoto), to finer texture (loam to clay loam) toward the western part of the Niger Delta region in the south. The inclusion of depth as a predictor variable significantly improved the prediction accuracy of RFM especially at lower depth intervals. These results could be used for producing soil function maps for national agricultural planning and in assessments of environmental sustainability.

Abbreviations: AfSIS, African soil information service; ALR, additive log-ratio; COK, compositional ordinary kriging; DEM, digital elevation model; DSM, digital soil mapping; EVI, enhanced vegetation index; GLM, generalized linear model; mALR, modified additive log-ratio; ME, mean error; MLR, multiple linear regression; MODIS, moderate resolution imaging spectroradiometer; NASA, national aeronautics and space administration; NDVI, normalized difference vegetation index; PSFs, particle size fractions; PTF, pedo-transfer function; RFM, random forest model; RK, regression kriging; RMSE, root mean square error; RTM, regression tree model; SRTM; shuttle radar topography mission; USDA, United State department of agriculture.

3.1 Introduction

There is a growing need for spatially continuous and quantitative soil information for environmental modelling and management (Minasny et al., 2008), especially at the national and supranational scale. Soil information is essential for global environmental challenges including climate change, food and water shortage, land degradation, and loss of biodiversity (Hartemink and McBratney, 2008). Such information is not always available at the required scale and coverage and in the right format (McBratney et al., 2003; Greve et al., 2012a). This is of concern in data-scarce countries where efforts for adequate soil data collection are often hampered by economic and logistic constraints.

The texture of the soil is one of its most important characteristics. It strongly affects water and nutrient retention, infiltration, drainage, aeration, SOC content, pH buffering and porosity and that affects many soil functions and mechanical properties. Soil texture is used at all levels in classification systems and in Soil Taxonomy it distinguishes soil orders (e.g. Vertisols or Alfisols) and is used all the way to the family level of particle size classes (Soil Survey Staff, 2010). Soil texture is used in the diagnosis of some key epipedons but particularly for argillic, natric, kandic horizons (Bockheim and Hartemink, 2013) Soil texture also determines the suitability of the soil for a particular use and management, waste disposal, and water management (Thompson et al., 2012). The capacity of soils to maintain organic carbon is influenced by its clay and silt content (Hassink, 1997; Bationo et al., 2007). PSFs are inputs in most hydrological, ecological, climatic and environmental risk assessment models (Ließ et al., 2012). The proportions of clay and sand particles have been used to create pedotransfer functions to estimate difficult-to-measure soil properties such as bulk density, hydraulic conductivity, water holding capacity, among others (Minasny and Hartemink, 2011).

Despite the importance, there is a dearth of information on soil texture (Scull et al., 2005), especially the PSFs at the resolution required for environmental modelling. In modeling, quantitative and continuous soil attributes rather than taxonomic soil classes are required (Gessler et al., 1996). However, most soil maps are produced as discrete class surface maps without considering the continuous variability of soil attributes with depth (Adhikari et al., 2013) and across space. Consequently, such soil maps lack quantitative information about the spatial distribution of very important soil physical attributes as required for effective environmental modelling, monitoring and management (Scull et al., 2005).

Digital soil mapping (DSM) offers a promising approach to spatial prediction of soil attributes. McBratney et al. (2003) formalized DSM in the now widely used *scorpan* model in which *S*, a set of soil attributes (*Sa*) or classes (*Sc*), is considered a function of other known soil attributes or classes (*s*), climate (*c*), organisms (*o*), relief (*r*), parent materials (*p*), age or time (*a*), and spatial location or position (*n*). All digital soil mapping (DSM) techniques involve establishing a relationship between the soil and environmental variables (representing the various soil forming factors) based on statistical and geostatistical models. Prediction is made at unobserved locations using the environmental variables at those locations and a soil property can be predicted using its interrelationships with the environmental covariates such as digital elevation models (DEMs) (McBratney et al., 2000), remotely sensed data (Odeh and McBratney, 2000), chemical and physical attributes obtained through laboratory analysis of soil sample or from legacy soil maps (Mayr, 2008).

Several applications of DSM techniques for predicting soil properties especially PSFs, using various statistical models, have been reported (Scull et al., 2005; Bishop and Minasny, 2006; Odeh et al.; 1995, 2007; Buchanan, et al., 2012; Greve et al., 2012b and Ließ et al., 2012). The models used in these studies are often based on compositional ordinary kriging (COK), regression kriging (RK), multiple linear regression (MLR), generalized linear model (GLM), regression tree model (RTM) and recently Random Forests (RF) with varying scale from the field to national level. However, very few DSM studies have been conducted in Sub-Sahara Africa where there is an urgent need for up to date spatial soil information (Sanchez et al. 2009). The objective of this study is to produce a fine resolution digital soil particle-size fractions map for Nigeria. We have used existing soil information (legacy data) and the latest DSM technologies to predict PSFs across the whole country.

3.2 Materials and methods

3.2.1 Study Area

Nigeria is located within latitudes 4° and 14° North, and longitudes 2° and 15° East, with a total area of about 923,768 km^2. The climate is humid in the south and semi-arid in the north. Seasonal rainfall distribution varies from 500 m to 4000 mm yr^{-1} with unimodal pattern in high rainfall areas close to the equator, low rainfall areas in the north, and bi-modal rainfall of between 1250 and 1500 mm (FAO, 1984). Temperatures throughout the year are in the range of 22-33^{o}C and rarely below 18^{o}C in any month. Vegetation ranges from evergreen forest in the southern part through moist Guinea savannas in the centre to the Sahel savanna in the northeastern part of the country.

Nigeria is comprised of inselbergs and sediments-filled basins derived through cycles of erosion from the cretaceous to the Pleistocene periods (Ojanuga, 2006). The country can be divided into highland and lowland areas (Iloeje, 2001). The highlands extend from the Jos plateau in the centre to the eastern border and the hills in some parts of the west. The lowlands are in the central part northward and southwards through Niger and Benue rivers and the coastal border (Udo, 1970). Nigeria is overlain by the Precambrian basement complex rocks mainly of igneous origin and sedimentary formations of Upper Cretaceous to recent age (Adejumo et al., 2012).

The major soils are Alfisols, Entisols, Ultisols, Inceptisols, Oxisols and Vertisols, (Soil Survey Staff, 2006; FDALR, 1990). According to FDALR (1990) Entisols dominates the soils of both the northern and southern fringes of the country with mostly Psamment and Aquent suborders. The central part of Nigeria is predominantly Alfisols and Ultisols with Ustalfs and Udults dominating the suborders. Nigeria has about 80 million hectares of arable land, of which 32 million hectares are cultivated. Major crops produced include beans, sesame, cashew nuts, cassava, cocoa , groundnuts, kolanut, maize, millet, palm tree, plantains, rice, rubber, sorghum, soybeans and yams.

3.2.2 Data processing

3.2.2.1 Soil data

Legacy soil profile data with PSF were obtained from the Africa Soil Profiles Database that was collated from reports of many decades of soil surveys and research conducted in Nigeria (Odeh et al., 2012; Leenaars, 2012). The data in the Africa Soil Profiles Database are from different periods. As soil texture is not a rapidly changing property compared to for example pH or SOC, we have not taken into account the year when the samples were taken. In the Nigerian soil survey reports the data are presented separately from genetic horizons for each profile. The samples were air-dried at room temperature, passed through a 2-mm sieve, and the fine-earth material was analyzed for PSFs using hydrometer and pipette methods. A number of particle-size fractions (coarse sand, fine sand, coarse silt, fine silt, sand, silt and clay) have also been reported by different soil surveyors. The size fractions were standardized into three fractions: clay (<2 µm), silt (2-50 µm) and sand 50-2000 µm) The PSFs were converted to g kg^{-1} as specified by GlobalSoilMap (Arrouays et al. 2014). In total the soil textural data from 978 soil profiles and 4568 layers (Table 3.1).was used in this study

Table 3.1 Summary statistics of Particle-size fraction profile data

Attribute	**Profile**	**Layers**	**Min**	**Max**	**Mean**	**SD**
			ISRIC Database			
Sand, %	1120	5034	0.0	100.0	58.0	24.0
Silt, %	1120	5034	0.0	89.0	16.3	12.3
Clay, %	1120	5034	0.0	88.1	25.4	18.6
			This study			
Sand, %	978	4568	0.0	100.0	57.9	24.3
Silt, %	978	4568	0.0	80.0	16.8	12.3
Clay, %	978	4568	0.0	88.1	25.3	18.8
Soil depth, cm	978		30.5	440.0	155.1	47.1

†Min, minimum; Max, maximum; SD, standard deviation

3.2.2.2 Fitting of mass-preserving profile spline function

In environmental modelling soil information is required at specified depth ranges rather than pedogenetic horizons. In this study, we fitted mass-preserving splines (Bishop et al., 1999) to the legacy soil profiles (n=978) to generate continuous PSF data at standard depth intervals (0-5, 5-15, 15-30, 30-60, 60-100, 100-200 cm) as following the GlobalSoilMap specifications (Arrouays et al., 2014). From the fitted splines of the raw data, the mean value of each PSF was derived for the six depths.

3.2.2.3 Additive log-ratio transformation

For the spatial prediction of compositional data, such as PSF, components have to sum to a constant, with distributions that are curtailed at the limits of 0 and 100 (De Gruijter et al. (1997). Therefore composition data must meet the following criteria:

d. Each of the components of the composition must be non-negative

$$Z * ij(x) \geq 0 \qquad [3.1]$$

where $Z * ij(x)$ is the estimate of a compositional regionalized variable, of jth component at ith location.

e. At each location, the components must sum to a constant

$$\sum_{j=1}^{n} Z * ij(x) = \emptyset, \text{ and } \emptyset\text{=constant.} \qquad [3.2]$$

f. Estimates of the composition should be unbiased

$$Z^*_{ij}(x) = \sum_{i=1}^{n} \lambda_i Z_{ij}(x) \ \sum_{i=1}^{n} \lambda_i = 1; j = 1, \dots, k. \qquad [3.3]$$

Out of k components in the composition, $Z^*_{ij}(x)$ represents the estimate of a compositional regionalized variable, of the jth component at the ith location.
For the interpretation of regionalized compositions the sample space is a positive (S^d) and not a multidimensional space (R^d) (Aitchison, 1990). A d-part simplex is thus defined as:

$$S^d = \left\{x = [x_1, x_2, \dots\dots, x_d];\ x_i > 0, i = 1,2, \dots\dots\dots, d;\ \sum_{i=1}^{d} x_i = k\right\} \qquad [3.4]$$

where S^d represent row vectors of d-part compositions; k is a constant, the sum of vectorial compositions which could be 100 (if composition is a percentage) or 1. The additive log-ratio (ALR), which allows transformation of the simplex S^d to the real space R^d, is expressed as

$$y_{ij}(x) = \ln\frac{z_{ij}(x)}{z_{ik}(x)} \quad k = d + 1 \quad i = 1, \dots, n \qquad [3.5]$$

where y_{ij} is the log ratio transformation of z_{ij}.

The inverse transformation of the above equation is:

$$z_{ij}(x) = \frac{\exp y_{ij}(x)}{\sum_{j=1}^{k} \exp y_k(x)} \qquad [3.6]$$

The results are that closure effect is removed, and subsequently through perturbation, the transformed data may fit a normal distribution, making the data suited to classical analysis such as MLR (Odeh et al., 2003).

We implemented a modified additive log-ratio (mALR) transformation (Odeh et al., 2003) of the spline-fitted PSFS dataset in R environment using the alr function of the compositions package (van den Boogaart and Tolosana-Delgado, 2008). Before this, a value of 0.001 was added to the three PSFs at each standard depth to remove the effect of zero values. The output of the transformation was two ALR-transformed variables (clay and sand) which were then used for predictive modelling. The predicted variates were later back-transformed using ALRInv function to three size fractions (clay, sand and silt) which were then used to determine the soil textural classes.

3.2.2.4 Environmental covariates

DEM tiles were obtained from the NASA SRTM data and mosaicked using the ArcGIS10 Data Management Toolbox. First and second derivatives like slope, aspect, curvatures (profile and plan), flow accumulation and compound topographic indices such as wetness index, stream power index were derived from the DEM using the ArcGIS10 Geomorphometry Toolbox (Reuter and Nelson, 2009). Landform classifications (Iwahashi and Hammond) based on algorithms developed by Iwahashi and Pike (2007) and Dikau et al. (1991) were also derived for Nigeria. Other covariates used were: global physiographic regions clipped for Nigeria (similar to Iwahashi-Pike landform), land cover map for year 2009, enhanced vegetation index (derived by Tor from MODIS on Terra), bands 1, 2, 3, 4 and 7 of Landsat 7-ETM+ coverage of Nigeria (obtained from Landsat GeoCover ETM+ 2000 edition) as well as digitized generalized geology and soil type maps of Nigeria. Average annual temperature and precipitation were interpolated from the 8-km grid coverage. The Normalized difference Vegetation Index (NDVI) was obtained from AfSIS website and clipped for Nigeria (AfSIS, 2012).

All the data layers were brought to the same projection and resampled to 1,000 m resolution using the nearest neighbour technique in ArcGIS10 Sample Toolbox. A total of 23 predictor variables were used in this study (Table 3.2). The environmental covariates were intersected to the six depths from the spline function. The dataset was randomly split into two sets: 67% for calibration (n=655) and 33% for validation (n=323). Prior to splitting, the entire dataset were first subset into the six geographical zones (northcentral, northeast, northwest, southeast, southsouth, southwest) and then combined to ensure uniform distribution of calibration and validation datasets.

Table 3.2 Description of environmental covariates

Variables	**Data source**	**Original scale and resolution**	**Type**	**Mean (Range)**
Slope Aspect	DEM	90 m	Q	180 (0-360)
Slope gradient	DEM	90 m	Q	1.11 (0-37.6)
Elevation	DEM	90 m	Q	328 (0-2360.8)
Wetness index	DEM	90 m	Q	6.26 (0.3-26.7)
Stream power index	DEM	90 m	Q	3.10 (-18-7.38)
Flow Accumulation	DEM	90 m	Q	43 (0-78402)
Plan curvature	DEM	90 m	Q	4.2x104 (-0.18-0.27)
Profile curvature	DEM	90 m	Q	4.2x104 (-0.23-0.21)
NDVI	MODIS vegetation indices	500m	Q	4386.5 (-1607-8225)
EVI	MODIS vegetation indices	500 m	Q	3147 (-967-6193)
Band 1	Landsat	30 m	Q	1289.5 (133-4463)
Band 2	Landsat	30 m	Q	3071.2 (117-5555)
Band 3	Landsat	30 m	Q	762.2 (1-4749)
Band 4	Landsat	30 m	Q	137.4 (0-255)
Band 7	Landsat	30 m	Q	98.3 (0-255)
Precipitation	Interpolated long-term mean precipitation	8 km	Q	1120 (272-2746)
Temperature	Interpolated long-term mean temperature	8 km	Q	268.2 (206-291.6)
Physiographic region	DEM	90 m	C	7 classes
Iwahashi	DEM	90 m	C	16 classes
Hammond	DEM	90 m	C	13 classes
Geology	Scanned and digitized geological map	1:5,000,000	C	14 classes
Soil types	FDALR	1:650,000	C	58 classes
Landuse	MODIS land cover map	500 m	C	15 classes

†C, categorical; DEM, digital elevation model; EVI, enhanced vegetation index; FDALR, Federal department of agriculture and land resources; NDVI, normalized difference vegetation index; Q, quantitative.

3.2.3 Spatial prediction of PSFs

3.2.3.1 Random Forest Model

Random Forest model (RFM) is developed as an extension of regression tree model (RTM) to improve the prediction accuracy (Breiman 2001a; Liaw and Wiener, 2002) and reduce model over-fitting. It is an assemblage of a number of classification and regression trees using two levels of randomization for every tree in the forest (Breiman, 2001b). RFM has advantages over many other prediction models because it is insensitive to noise or weak prediction variables as it selects the most important variable at each node split (Okun and Priisalu, 2007), has reasonable prediction performance even with noisy predictor variables (Diaz-Uriarte and de Andres, 2006), and insensitive to missing values or outliers in a given dataset. In this study we employed the randomForest 4.6 package (Liaw and M. Wiener, 2002) in R environment to predict the PSFs. The Random Forest regression algorithm can be described following Liaw and Wiener (2002) and Hastie et al. (2009):

1. For j=1,…., n; draw a bootstrap sample Z^* of size n_{tree} from the original data then
2. Grow a random-forest tree $T_j$to Z, by recursively repeating the following steps at each terminal node of the tree, until the minimum node size $n_{min}$is reached.

a. Select M_{tree} variables at random from the predictors,p.
b. Choose the best variable at random/split-point amongm.
c. Split the node into two daughter nodes but before each split, select $m \leq p$ of the input variables at random as candidates for splitting.

3. Finally, output the ensemble of trees$\{T_j\}_1^n$ and predict new data by averaging the predictions of the n_{tree} trees

The RFM regression prediction at a new point x after n trees $\{T(x;\ \theta_j)\}_1^n$ are grown is expressed as:

$$\hat{f}_{rf}^{n}(x) = \frac{1}{n}\sum_{j=1}^{n} T(x;\ \theta_j) \qquad [3.7]$$

where $\theta_j$describes the jth random forest tree at each node and terminal-node values in terms of split variables.

Three parameters control the fitting of Random Forest models: (i) the number of trees (n_{tree}), (ii) the minimum number of samples in the terminal node n_{min}, and (iii) and the number of predictors to be used for the fitting of each tree (M_{try}) (Grimm et al., 2008). The M_{try} is a crucial parameter as it defines the strength of each individual tree and the correlation between any two tree in the RF model. Normally for regression, the default value for M_{try} is $p/3$ and n_{min} is 5 (Hastie et al., 2009). We used the "train" function of the "caret" R package to determine optimum M_{try}value for modelling at each depth interval. The "train" function tunes various models by selecting a combination of sensitive parameters that are associated with the optimal resampling statistics of held-out samples. These are used to fit the final model with the entire training dataset. The relative importance of the predictor variables in modelling PSFs for Nigeria was assessed using the "importance" function in the "randomForest" R package.

3.2.3.2 Soil sampling depth as a predictor variable

The inclusion of soil sampling depth as a predictor variable to estimate soil properties (especially bulk density) by pedo-transfer functions (PTFs) is well-established (Tamminen & Starr, 1994; et al., 2010; Minasny and Hartemink, 2011). Here we evaluate the contribution of sampling depth in modelling PSFs. We first added vectors of mean of the various depth intervals (e.g. 2.5, 5, 10, 22.5, 45, 80 and 150 for 0-5, 5-15, 15-30, 30-60, 60-100, and 100-200) to the set of predictor variables for each depth. Thereafter we stacked datasets of the six depth intervals together to obtain one single data set which was used to fit a single model. The predicted values were subset into the six standard depths for model assessment and validation. The single model produced was then used to make predictions onto the entire study area at the six different depths. To account for soil sampling depth in the grids, the mean value of each depth interval (e.g 2.5 for 0-5 cm depth) was populated in the grid used in predicting for that particular depth.

3.2.3.3 Model Accuracy

To evaluate the prediction performance by the three models, we divided the dataset into two separate subsets by a random selection process using the sample function in R prior to modelling. In using the sample function, approximately 2/3 (n=655) were earmarked for model calibration while the remaining 1/3 (n=323) was used for cross validation. The following four parameters were computed on the validation subset, using the R statistical software package (R Core Development, 2013).

a. Coefficient of determination (R^2) a measure of the percentage of variation explained by each model:

$$R^2 = \frac{\sum_{i=1}^{n}(p_i - \bar{o}_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o}_i)^2} \qquad [3.8]$$

where n denotes data points, o_i and p_i are observed and predicted PSFs values at the i^{th} point, $\bar{o}_i$ and $\bar{p}_i$their respective means, respectively.

b. Mean error (ME) a measure of model's prediction bias:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(o_i - p_i) \qquad [3.9]$$

c. Root-mean-squared-error (RMSE) a measure of model accuracy:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \quad [3.10]$$

d. Lin's concordance correlation coefficient (CCC), a measure of the strength of the agreement between the observed and predicted PSF values:

$$\rho_c = \frac{2\rho\sigma_o\sigma_p}{\sigma_o^2+\sigma_p^2+(\mu_o-\mu_p)^2} \quad [3.11]$$

where ρ_c is the estimated CCC, μ_o and μ_p are the means for the raw and predicted PSFs while σ_o^2 and σ_p^2 are the corresponding variance and ρ the Pearson correlation coefficient between the raw and predicted PSFs. Generally, a good model will have R^2 and ρ_c close to 1 and ME and RMSE close to 0.

3.3 Results and discussion

3.3.1 Soil Legacy data

3.3.1.1 Spatial distribution of soil profiles and covariates

The spatial distribution pattern of the 978 soil profiles used in this study is presented in Fig. 3.1. There is fair spread of the soil profiles across the country but some areas in the northeast (NE) and southwest (SW) have higher density of sampled profiles. In the northcentral and southeast there were fewer pedons. Most of the early soil survey projects in Nigeria were guided by interests in food and cash crop production (Odeh et al., 2012) regions with high production capacity for cash crops (e.g cocoa in the SW) and food crops (cereal grains in the NE) were densely surveyed and sampled. Also, areas with agricultural research institutions have larger number of pedons such as the Jos-Kaduna-Zaria axis hosting the Institute for Agricultural Research (IAR) and areas around Ibadan that host of the International Institute of Tropical Agriculture (IITA) and the Institute for Agricultural Research and Training (IAR&T).

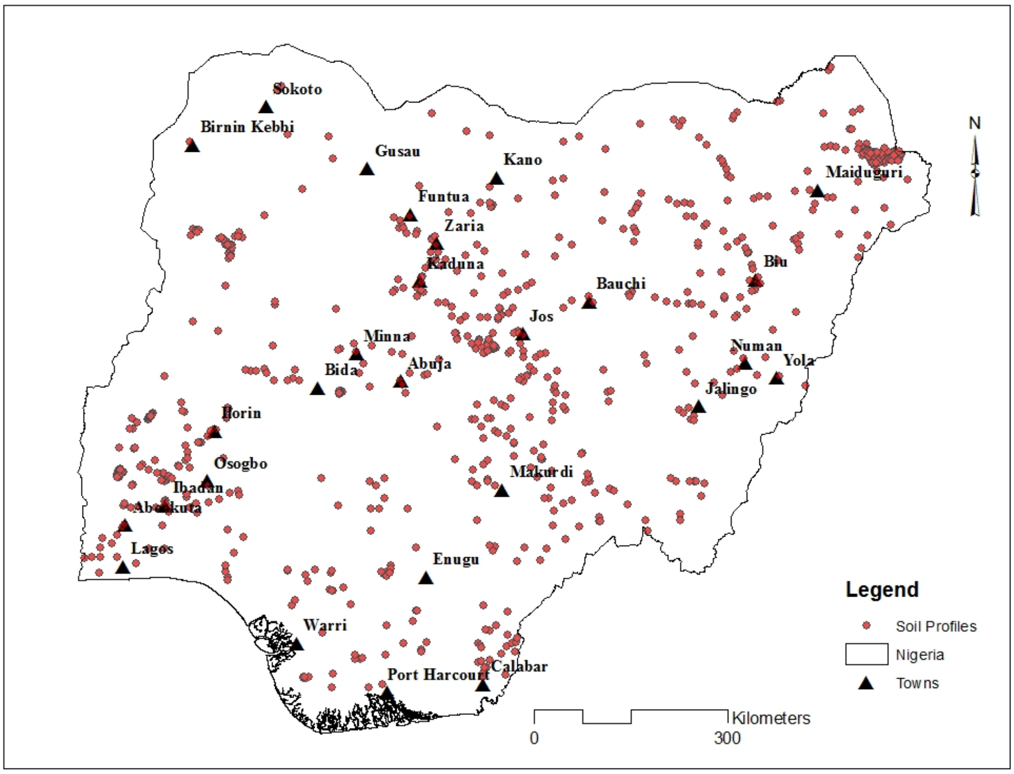


Figure 3.1 The distribution of legacy soil profile (n=978) in Nigeria.

The distribution patterns of environmental covariates in the grid and sample locations are comparable around the mean, but the extreme values of most covariates were not well covered by the legacy profile points (Table 3.3). This could affect validity of our predictions as we are most likely predicting outside the range of values upon which the model was built. As shown in Fig. 3.1, future sampling in areas of sparse data like the Sokoto axis of the northwestern region, as well as the northcentral and southeastern region, could help overcome this challenge.

Table 3.3 Summary statistics of environmental covariates at legacy soil profile locations and grids

	Elevation		**Aspect**		**EVI**		**Rainfall**		**Slope**	
	Sample	**Grid**	**Sample**	**Grid**	**Sample**	**Grid**	**Sample**	**Grid**	**Sample**	**Grid**
Min	6.7	0	0.5	0	556	-943	286.1	272.2	0	0
Q1	186.2	177	82.8	90.2	2426	2406	764.5	767.8	0.2	0.3
Median	289.2	304.8	182.4	179.6	3046	3076	1159.8	1108.3	0.7	0.6
Mean	338.7	328.6	179.7	180.7	3081	3091	1085.3	1120.8	0.9	1.1
Q3	390.8	428.7	275.9	272.7	3715	3627	1238.9	1285.8	1.1	1.1
Max	1371.7	2279.2	359.8	360	5291	5776	2692.2	2743.6	21	35.9

†Max, maximum; Min, Minimum; Q1, first quartile; Q3, second quartile

3.3.1.2 Summary statistics of spline-fitted particle-size fractions at continuous depth intervals.

A summary of the predicted PSFs from the equal-area quadratic splines is presented in Table 3.4. The frequency distributions of the PSF data are typical given that clay and silt are positively skewed whereas sand is skewed slightly negative. Similar observations have been reported elsewhere (Adhikari et al., 2013). The sand fraction has a higher variation (SD 22 to 26 gkg^{-1}) compared to clay (SD 16 to 19 $g\ kg^{-1}$) and silt (11 to 14 $g\ kg^{-1}$) as was found by Buchanan et al. (2012) and Adhikari et al. (2013) but it is in contrast to Odeh et al. (2003) and Oku et al. (2010) who reported a higher variability in clay content compared to sand and silt. The variation is relatively high but considering the heterogeneity of the landscape as well as the large extent of our study area, such is expected.

The clay content increases from the top 30 cm depth with a peak at the 60-100 cm likely caused by clay illuviation (Osei and Okusami, 1994; Ayuba et al., 2007; Sharu et al., 2013). The increase in clay content (>20%) with depth is diagnostic of the major soil types (Alfisols, Ultisols) in Nigeria (Osei and Okusami, 1994; Amhakhian and Achimugu, 2011). The mean sand content is higher than the clay and silt contents for each depth which is commonly found in the soils of West Africa (with exception of Vertisols) (Jones and Wild, 1975).

3.3.2 Performance of RFM in predicting particle-size fractions

The model performance parameters (Eq. [3.8]-[3.11]), were used to assess the quality of prediction of PSFs (Table 3.5). Results showed that the combination of the various predictor variables can explain 16 to 53%, 21 to 48% and 21 to 26% of the variation in clay, sand and silt contents respectively. This is within the range reported for clay and sand contents in other studies using similar prediction models (Ließ et al., 2012; Adhikari et al., 2013) but outside the range reported for silt content. In Nigeria, low R^2 has been reported for prediction of silt content (Ugbaje and Reuter, 2013) but our predictions show an improvement over their report. This could be attributed to the effect of ALR as Odeh et al (2003) reported an improved prediction accuracy when PSFs are transformed using ALR before fitting predictive models.

The model performed significantly better at the top 30 cm (0-5, 5-15 and 15-30 cm) compared to the lower layers (30-60, 60-100 and 10-200 cm). Similar results have been reported by several others (Henderson et al., 2005; Minasny et al., 2006; Malone et al., 2009; Vasques et al., 2010; Kempen et al., 2011; Adhikari et al., 2013; Ugbaje and Reuter, 2013). This could be attributed to the nature of the environmental variables used (Adhikari et al., 2013) and effect of lower data density with depth. Most environmental covariates used in this study are based on land surface characteristics and are likely to have stronger relationship with topsoil than subsoil properties. The prediction performance for the lower depths could be improved by inclusion of covariates such as Gamma-radiometric (K, Th, U) or electromagnetic induction (EM) (Cooke, 1996; Rawlins et al., 2009; Priori et al., 2014). However considering the extent of Nigeria the cost of acquiring this data may be too exorbitant to off the extra benefit.

Table 3.4 Summary statistics of spline-fitted particle-size fractions (%) at six standard depth intervals.

PSF	Statistics	Depth(cm)					
		0-5	5-15	15-30	30-60	60-100	100-200
Clay	Max	84.3	84.2	87.6	88.2	86.5	85.7
	Min	0	0	0	0	0	0
	Mean	19.0	20.0	22.9	27.7	29.9	27.8
	SD	19.39	19.22	19.21	19.18	17.46	16.04
	SEM	0.621	0.615	0.615	0.614	0.559	0.513
	Skewness	1.63	1.54	1.25	0.80	0.52	0.47
	Kurtosis	4.88	4.65	3.91	2.98	2.72	2.98
Sand	Max	100	100	98	97.3	99.1	100
	Min	0.7	1	0.5	0.2	0.2	0.2
	Mean	62.4	61.7	59.4	55.2	53.5	56.2
	SD	25.56	25.43	25.32	24.62	22.8	21.56
	SEM	0.818	0.814	0.81	0.788	0.73	0.69
	Skewness	-0.75	-0.7	-0.57	-0.38	-0.23	-0.15
	Kurtosis	2.37	2.28	2.12	2.03	2.12	2.36
Silt	Max	79.5	77.5	74.4	70.7	64.4	59.1
	Min	0	0	0	0	0	0
	Mean	18.5	18.3	17.7	17.1	16.6	16.0
	SD	13.6	13.26	12.65	11.65	11.12	10.97
	SEM	0.435	0.424	0.405	0.373	0.356	0.351
	Skewness	1.08	1.07	1.09	1.07	0.88	0.77
	Kurtosis	3.93	3.94	4.07	3.99	3.33	3.07

†SD, standard deviation; SEM, standard error of mean; Max, maximum; Min, Minimum.

In terms of prediction accuracy, sand content had the highest RMSE values across all depths whereas the lowest RMSE was associated with the prediction of silt at all depth intervals. This trend corroborates the reports of other studies using similar modelling approaches (Buchanan et al., 2012; Niang et al., 2013) but slightly different from the report of Odeh et a.l (2003). The lower RMSE of the silt content in this study is expected since silt was not used in modelling and it was a product of the back-transformation of the initial ALR variates.

RF_d significantly improved the model performance especially at the lower depths (Table 3.5). With RF_d there is similar model performance for the various PSFs. The inclusion of soil sampling depth improves the performance of RFM by 67-100% (R^2 values). This supports the inclusion of soil depth as a predictor variable to improve prediction of soil attributes.

3.3.3 Predictor variables for predicting soil particle-size fractions

A key advantage of RFM in comparison with classical multiple regression models is that the latter involves feature selection through stepwise and criterion-based procedures in which one or two of the highly correlated predictor variables are typically retained with the rest discarded. In contrast, RFM “spreads” the importance of predictors in the model across all the predictor variables (Cutler et al., 2007). RFM estimates the relative importance of the predictor variables, based on how worse the prediction would be if the data for a particular variable were permuted randomly (Prasad et al., 2006). This approach guards against the elimination of good predictors variables which may be pedologically important, although are highly correlated with each other. We used the “importance” function in the “randomForest” package to assess the importance of predictor variables used to predict PSFs.

The predictor variables showed a varying level of importance in the model (Fig. 3.2). There was a large influence of climatic elements (precipitation, temperature), vegetative indices (EVI, NDVI), terrain attributes (elevation, stream power index and slope), soil types, geology and Landsat bands on the spatial distribution of PSFs. However, the relative importance of these variables varies with depth and from one fraction to another. Other studies have also reported the relationship between terrain attributes and soil properties especially PSFs (Moore et al., 1993; Odeh et al., 1995; Thompson et al., 2006; Greve et al., 2012a; 2012b; Ließ et al., 2012) with terrain attributes explaining between 20% and 88% of the variation in soil properties

(Thompson et al., 2006). This could be attributed to their impact on vertical and lateral movement of soil particles through erosion and disposition. In Nigeria the influence of geology and soil types on the spatial distribution soil texture has been documented in previous studies (Osei and Okusami, 1994; Law-Ogbomo and Nwachokor, 2010).

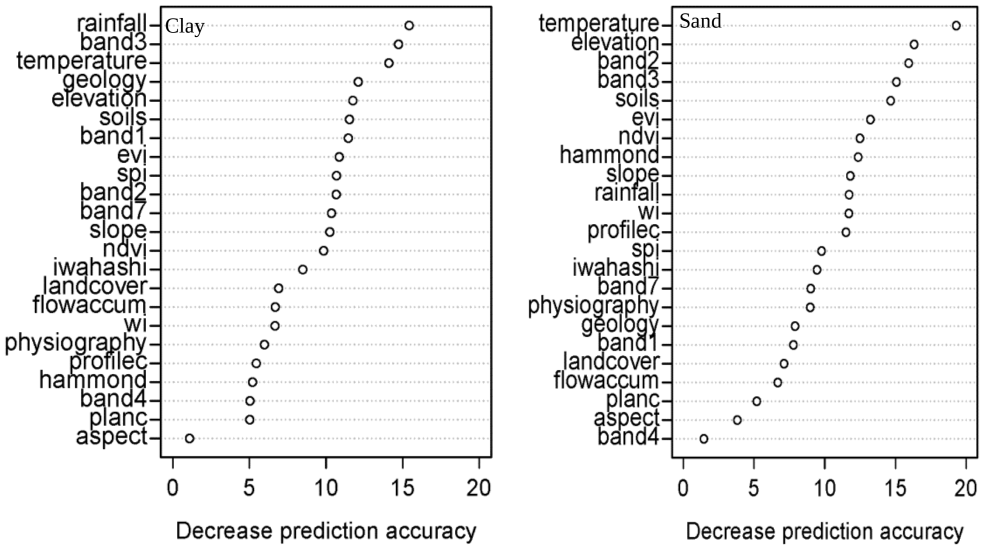


Figure 3.2 Illustration of variable importance derived from random forest models of soil particle size fractions for Nigeria. Abbreviations: evi; enhance vegetation index, spi; stream power index, ndvi; normalized difference vegetation index, wi; wetness index, profilec; profile curvature, planc; plan curvature.

Table 3.5 Performance of Random Forest and Random Forest with inclusion of depth as a predictor in modelling Particle size fractions

		RFM				**RF_d**			
PSF	Depth	ME	RMSE	R^2	pc	ME	RMSE	R^2	pc
	cm	%	%			%	%		
Clay	0-5	3.53	13.59	0.53	0.65	-0.4	6.48	0.89	0.94
	5-15	3.4	13.11	0.56	0.69	0.35	5.71	0.91	0.95
	15-30	2.95	13.38	0.54	0.68	2.26	6.93	0.89	0.94
	30-60	2.6	14.98	0.42	0.62	-0.23	10.04	0.72	0.85
	60-100	4.01	15.7	0.29	0.46	1.58	10.56	0.68	0.82
	100-200	4.21	15.6	0.16	0.3	1.18	12.63	0.43	0.64
Sand	0-5	-6.51	19.67	0.48	0.6	0.03	7.7	0.91	0.95
	5-15	-6.03	19.26	0.49	0.63	-0.42	7.05	0.92	0.96
	15-30	-5.26	18.79	0.49	0.63	-1.97	7.26	0.92	0.96
	30-60	-4.14	18.81	0.43	0.61	-0.74	9.55	0.85	0.92
	60-100	-5.71	19.48	0.33	0.5	-2.09	11.52	0.76	0.87
	100-200	-6.67	19.86	0.21	0.36	-1.91	15.85	0.51	0.7
Silt	0-5	2.99	12.22	0.26	0.39	0.37	4.44	0.88	0.94
	5-15	2.63	11.72	0.27	0.42	0.08	4.14	0.9	0.95
	15-30	2.31	10.96	0.25	0.39	-0.29	3.43	0.91	0.95
	30-60	1.54	9.82	0.24	0.41	0.97	5.17	0.82	0.89
	60-100	1.69	9.73	0.24	0.4	0.51	5.66	0.76	0.85
	100-200	2.46	10.06	0.21	0.35	0.74	7.15	0.59	0.74

†PSF; Particle size fractions, ME; Mean error, RF_d; Random Forest using soil depth as a predictor, RMSE; Root mean square error, p_c; Lin's concordance correlation coefficient.

3.3.4 Spatial prediction of particle-size fractions

The descriptive statistics of sand, silt, and clay fractions predicted for various depth intervals is presented in Table 3.6. The distribution of the predicted PSFs by RF and RF_d follow a similar pattern as the spline-fitted data. The RF_d slightly reduced mean values of predicted PSFs and predicted PSFs show lesser variability than the spline-fitted data. This could be attributed to the smoothening out of outliers as prediction models tend to have smoothening effect (Odeh et al., 1995). In addition, RF grows a large number of unpruned trees and makes final prediction using the average prediction of the entire trees and as such tends to overcome model overfitting that is common among prediction models.

Figs. 3.3-3.5 show the maps of predicted PSFs. There is an increase in the clay content with depth especially in the southern part of the country (see Fig. 3.3). The magnitude of this vertical increase in clay content differs (Fig 3.3). At some locations this is steady and gradual while it is abrupt in others; giving rise to a bulge of clay with depth. The gradual increase of clay content with depth has also been reported for Nigeria (Moberg and Esu, 1991; Olowolafe, 2002; Ayuba et al., 2007, Sharu et al., 2013). Fig. 3.3 reveals an increase of clay content with depth in the southern part of the country compared to the northern part. This supports the work of Vine (1987) who reported an increase in clay content with depth in soils of southern Nigeria except those in valley bottoms. This pattern is the result of vertical clay movement (eluviation/illuviation), faunal perturbation (Oyodele et al., 2006; Sharu et al., 2013) and movement of clay particles due to soil erosion (Amusan et al., 2005; Salako et al., 2006). According to Vine (1987) these factors affect pedogenetic processes through the incorporation of dust, in addition to mixing of coarser and finer layers of sediments.

Table 3.6 Summary statistics of predicted particle-size fractions.

PSF	Depth (cm)	Max	Min	Mean	RF SD	Skewness	Kurtosis	Max	Min	Mean	RF_d SD	Skewness	Kurtosis
Clay	0-5	70.4	0.4	15.5	5.77	1.74	7.17	99.2	0	15.8	10.37	2.29	7.78
	5-15	74.6	0.1	16.2	7.47	1.34	4.38	99.2	0	16.2	10.54	2.22	7.37
	15-30	79.3	2.0	20.8	6.92	1.32	4.01	99.1	0	18.1	11.16	1.93	5.69
	30-60	98.2	0.2	25.3	9.16	1.51	5.75	98.8	0	25.3	11.34	1.08	3.28
	60-100	74.9	0.6	29.8	8.73	-0.50	1.04	98.8	0	26.5	11.18	0.78	2.83
	100-200	65.2	1.4	28.4	7.69	-0.76	0.98	98.7	0	26.4	10.72	0.45	2.07
Sand	0-5	97.3	10.1	67.4	9.69	-0.92	2.08	100	0.4	66.9	15.45	-1.08	2.31
	5-15	99.7	6.3	67.1	12.67	-0.51	1.10	100	0.4	66.5	15.47	-1.07	2.28
	15-30	91.6	8.2	63.6	10.11	-1.02	1.75	100	0.4	64.9	15.70	-0.98	1.90
	30-60	91.2	1.4	58.9	11.01	-0.95	1.79	100	0.2	59.1	15.59	-0.51	1.36
	60-100	98.6	8.7	54.3	11.94	0.50	0.77	100	0.2	58.3	15.31	-0.26	1.26
	100-200	94.0	15.9	55.6	10.2	0.79	0.29	100	0.3	58.5	14.92	-0.02	0.87
Silt	0-5	46.3	0.6	17.2	5.38	0.32	0.29	75.3	0	17.4	7.87	0.51	1.09
	5-15	46.3	0	16.7	6.42	0.09	0.09	76.7	0	17.3	7.73	0.53	1.21
	15-30	42.0	0.6	15.6	4.79	0.45	0.52	79.3	0	16.9	7.38	0.53	1.29
	30-60	45.2	0.1	15.8	4.80	0.50	0.88	74.7	0	15.6	6.32	0.44	1.58
	60-100	39.6	0.1	15.9	4.73	-0.04	0.54	72.3	0	15.3	6.10	0.29	1.28
	100-200	37.2	0.7	16.0	4.50	-0.01	-0.23	48.6	0	15.1	6.01	0.10	0.60

†SD, standard deviation; Max, maximum; Min, Minimum; RF, Random Forest; RF_d; Random Forest using soil depth as a predictor.

There are also patches of high-to-medium clay content around the Lake Chad, Biu, Jos and Mambilla plateaus as well as the coastal Niger delta area which was also found by previous studies (Lombin and Esu, 1988; Moberg and Esu, 1991, Olowolafe, 2002). The Lake Chad and Niger-Delta areas receive colluvium materials and lacustrine deposits which explains the high to medium clay content. The prevalence of Quaternary volcanic rocks (basalt, lava flows and ash deposits) accounts for the high clay content around Jos and Mambilla plateaus (Olowolafe, 2002). The relatively high clay content in the subsurface layer of soils around Lagos and Enugu (Fig. 3.3) seems anomalous considering that these areas are overlain by Tertiary and Upper Cretaceous sandstones. The high clay content around these areas has been reported (Vine, 1987) and was attributed to sporadic clay beds in the sandstones which accumulated overtime while the sandy surface was gradually lost by soil erosion.

The sand content (Fig. 3.4) of soils in Nigeria is relatively high compared to clay and silt across the entire country. This can be attributed to variation in parent material and partly due to Aeolian deposition of sands from the Sahara desert. About 50% of Nigeria's landmass is underlain by sandstones of Cretaceous age (Adeleye and Dessauvagie, 1970; Hassan, 2010). According to Ogunwale et al (1975) soils derived from sandstones cover about 18% (160,000 km^2) of the surface area of Nigeria.

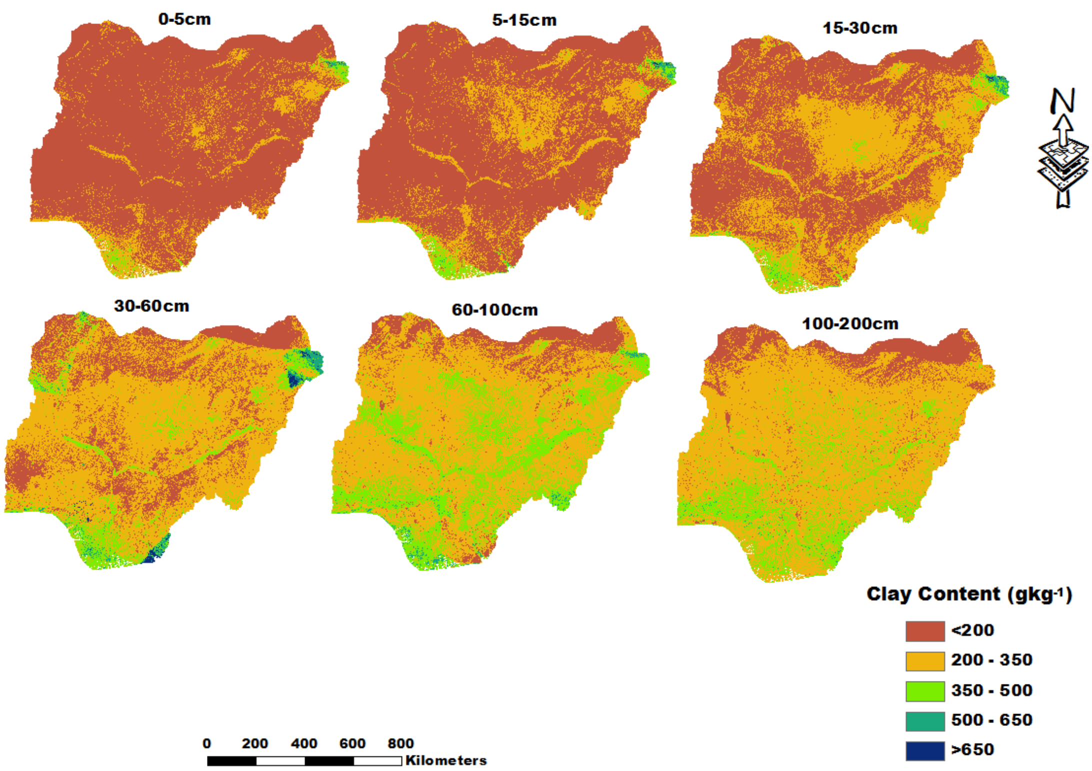


Figure 3.3 Spatial distribution of predicted clay content using random forest in Nigeria.

There is an area of soils with high-to-medium sand content at the border of northern Nigeria (Fig. 3.4) which is caused by deposition of sand from the Sahara desert (Sombroek and Zonneveld, 1971; McTainsh, 1984) during the Pleistocene when the Sahara desert extended further southward (Grove, 1958). Accompanying the southern extension of aridity is the building and migration of sand-dunes with deposition of windblown sands in the direction of north-east to south-west (Chartres, 1982). The sand content however decreased gradually southwards and with depth supporting the work of Omoregie (1998). The soils are moderately sandy in the south west and south-eastern part of the country that could be attributed to the coarse nature of the predominant parent materials in these regions. They are overlain by weathered sandstones of Palaeocene/Pleistocene age and gneiss of the Precambrian basement complex (Smyth and Montgometry, 1962; Igwe et al., 2009).

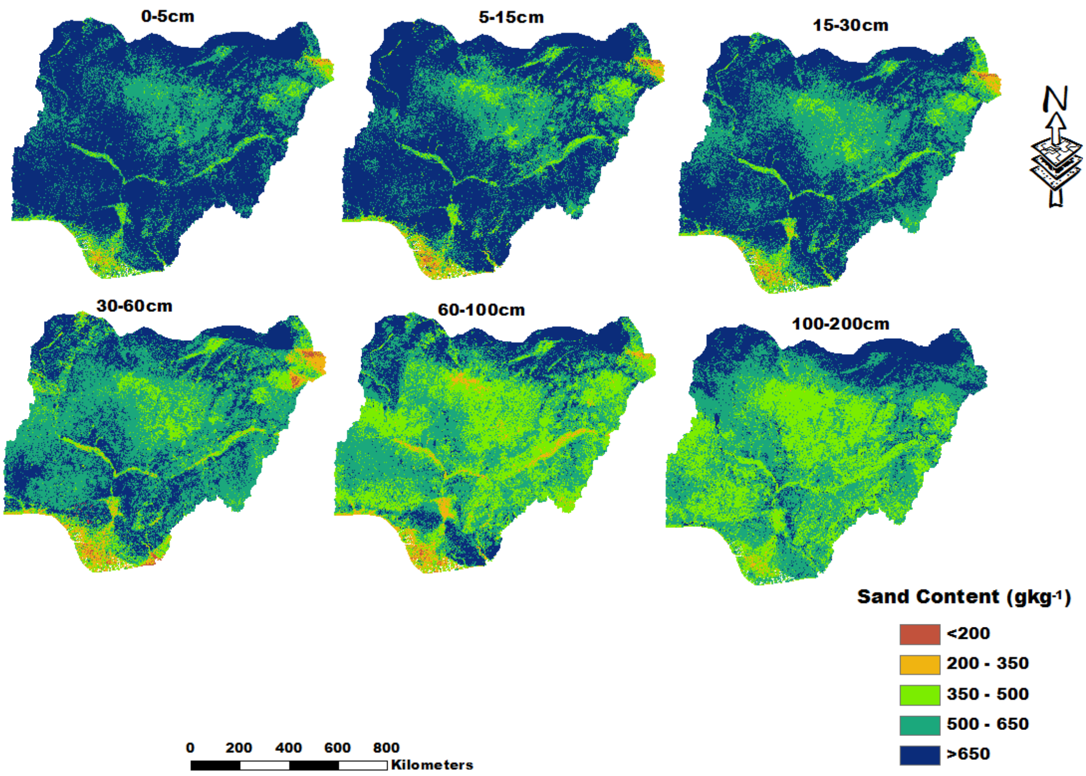


Figure 3.4 Spatial distribution of predicted sand content using random forest in Nigeria.

The silt content of the soils in Nigeria is relatively low (Fig. 3.5) and has been reported previously (Ojanuga, 1975; Igwe, 2005). However, soils with medium silt content occur around Zaria-Funtua-Kano axis and in the Niger-Delta areas as reported in previous studies (Bennett, 1980; Morberg and Esu, 1991). Most soils of the Kano plains are silty fine sands derived from wind-sorted desert sands (Lawes, 1962) or Aeolian drifts (Tomlinson, 1961; Higgins, 1963; Klinkenberg and Higgins, 1968). Maniyunda et al. (2013) reported high silt content in soils from Funtua and Katsina area. Relatively high silt content has also been reported for inland valley bottom soils in the coastal southern part of the country (Ogban and Babalola, 2003).

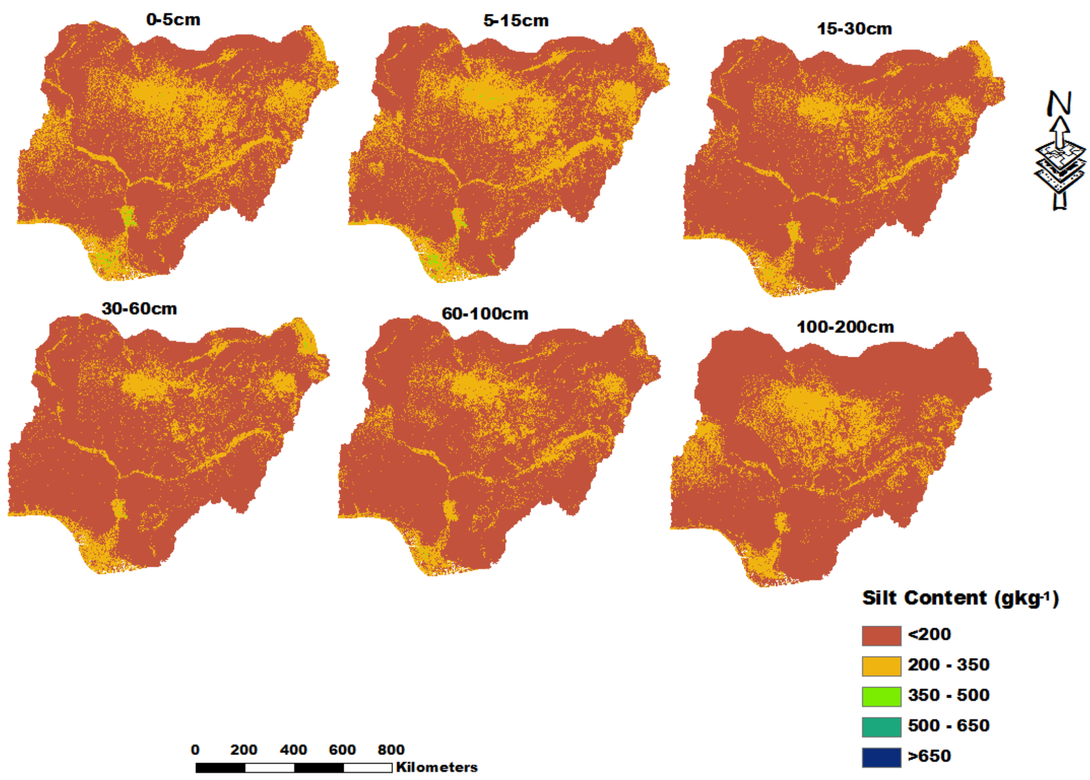


Figure 3.5 Spatial distribution of predicted silt content using random forest in Nigeria.

3.3.5 Spatial distribution of soil texture

We present here the patterns of soil texture for the six layers, as predicted by RFM (Fig. 3.6). The variation in soil texture shows a progressive transition from a coarse-texture (sand) along the fringes of northern Nigeria (e.g upper part of Maiduguri and Sokoto), to finer texture (loam to clay loam) towards the western part of the Niger delta in the south. The orientation of this transition in soil texture, especially the top 30 cm layers, suggests the direction of the prevailing north-easterly wind which deposit Aeolian sediments. Generally, the soils are mainly sandy-loam, loamy-sand, sandy clay-loam, clay-loam, sandy-clay and clayey with the soils becoming more clayey and less sandy with depth (Fig. 3.7). This is typical of the major soil types present in the study area: Alfisols, Ultisols, Oxisols and Entisols (Soil Survey Staff,

2006) as was reported by Igwe, 2005; Sharu et al., 2013; Maniyunda et al. (2013). On area basis, soil texture of Nigeria ranges from sand (4.2 x 10^6 ha) to sandy loam (5.3x10^7ha) in the surface layers and from sandy clay loam (5.2 x 10^7 ha) to clay (6.9 x 10^6 ha) in the subsoils respectively.

The general pattern of soil texture in Nigeria has been attributed to the influence of the combination of the differences in parent material (Akamigbo and Asadu, 1983), pedogenetic processes involving clay movement (Hassan, 2010), in addition to contributions from Aeolian dust (Vine, 1987; Morberg and Esu, 1991; Kparmwang, 1993). In Nigeria, parent materials vary from very coarse pegmatite to fine grained schist, and from acid quartzite to basic rocks consisting largely of amphibolites (Smyth and Montgomery, 1962; Hekstra and Andriesse, 1983). Law-Ogbomo and Nwachokor (2010) reported that soils developed on basalt exhibits fine texture (sandy clay loam to clay) with those from sandstone having medium texture (sandy loam to sandy clay loam) and soils from coastal plain sands very coarse texture (loamy sand to sand). They observed that soils developed on basement complex rock and shale exhibit similar textures ranging from loamy sand to sandy clay loam.

3.4 Conclusions

Developing DSM models by correlating soil and predictor variables is an efficient but challenging quantitative spatial prediction approach, especially is a situation with sparse soil profile data. This study provides an example where a geodatabase of important soil attributes can be populated from a limited soil dataset. We demonstrate the robustness of RFM to predict soil particle-size fraction as compositional data for Nigeria using legacy soil data. Considering the dearth of soil profile data used in this study the results presented here are a good first approximation of digital mapping of these soil attributes for Nigeria. No doubt, work will continue to improve on this first approximation as more data becomes available.

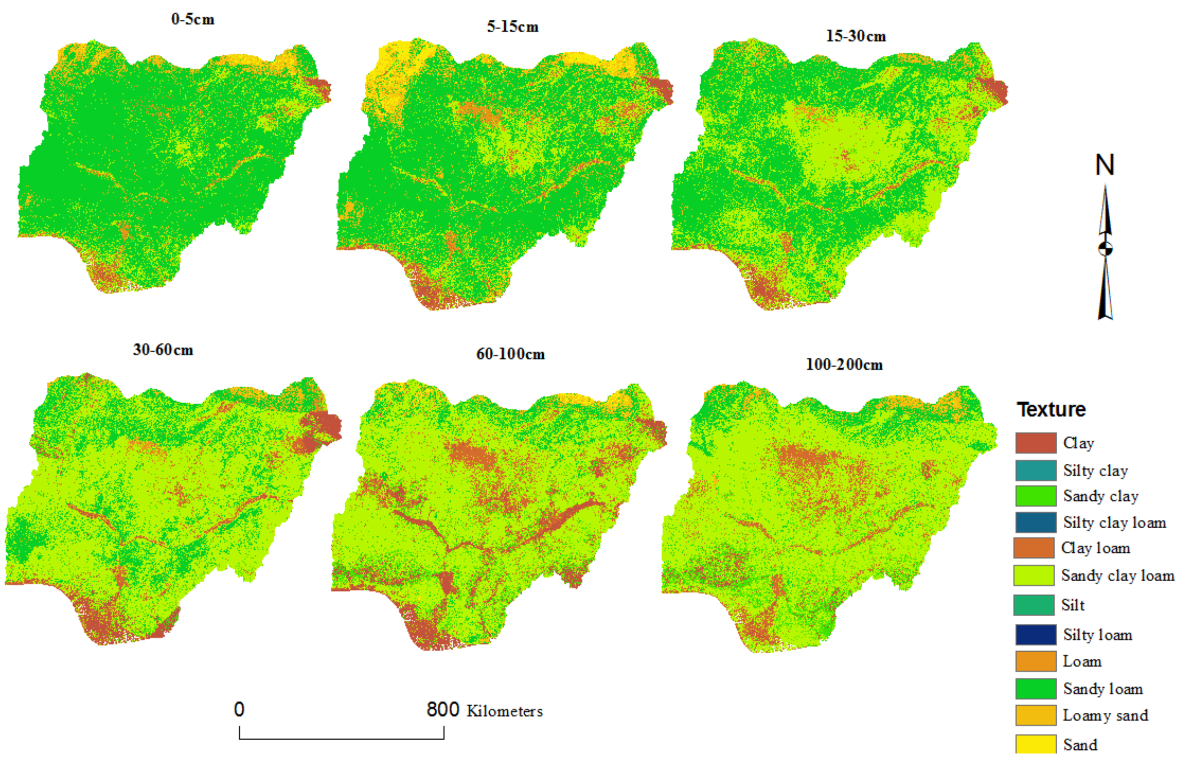


Figure 3.6 Spatial distribution of predicted soil texture using random forest in Nigeria.

Generally, from this study the following salient points are adduced:

1. Nigerian soils are predominantly coarse-textured with texture gradually becoming finer southwards; the Northern region of the country having a higher sand content.
2. Soil texture ranges from sand (4.2 x 10^6 ha) to sandy loam (5.3x10^7ha) in the surface layers and from sandy clay loam (5.2 x 10^7 ha) to clay (6.9 x 10^6 ha) in the subsoils
3. RFM is robust in predicting PSFs while the inclusion of soil depth as predictor significantly improved the model accuracy.
4. In modelling PSFs for Nigeria, terrain attributes (elevation, stream power index, and slope), soil types, vegetative indices, as well as climatic variables (especially precipitation and temperature) are the most important predictors.

These results could be used for producing soil function maps (e.g. water holding capacity) or for national agricultural irrigation planning and for assessing for environmental sustainability

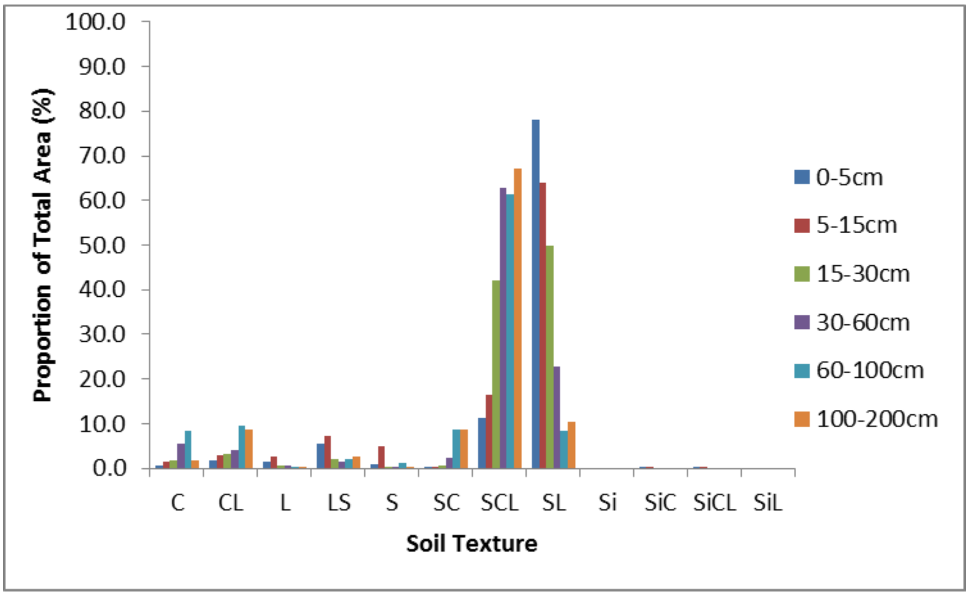


Figure 3.7 Predicted soil texture classes as a percentage of total area of Nigeria using RFM. Abbreviations: C, clay; CL, clay loam; L, loam; LS, loamy sand; SC, sandy clay; SCL, sandy clay loam; S, sand; SL, sandy loam; Si, silt; SiC, silty clay; SiCL, silty caly loam; SiL, silty loam.

Acknowledgements

The authors wish to thank the University of Sydney for providing a scholarship through her International Scholarships program and Ahmadu Bello University Zaria for given a study fellowship to the first author. We also acknowledge the efforts of Soil Scientists in Nigeria that carried out various soil surveys, which were the original sources of data used for this work; we thus thank ISRIC in the Netherlands and other institutions as custodians of the soil survey legacy data used in this study. The authors also appreciate the editorial comments by the editor and anonymous reviewers which went a long way to improving the quality of the paper.

References

Adejumo, T. W., Alhasan, M., & Boiko, I. L. 2012. Physico-mechanical properties of some major weak soils in Nigeria. *Electronic Journal of Geotechnical Engineering,* 17, 2435-2441.

Adeleye D.R. & Dessauvagie. T. 1970. Stratigraphy of Nigeria embayment near Bida, Nigeria. In: Dessaubvagie, T.F.J. and D.R. (Eds.). Conference of African Geology proceedings, University of Ibadan, Nigeria p 7-14.

Adhikari, K., Kheir, R. B., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A.B. & Greve, M.H. 2013. High-Resolution 3-D mapping of soil texture in Denmark. *Soil Science Society of America Journal*, 77(3), 860-876. doi: 2136/sssaj2012.0275.

Africa Soil Information Service (AfSIS). 2012. AfSIS MODIS Data Sets: Normalized Difference Vegetation Index (NDVI) Long Term and Monthly Averages. Palisades, NY: Center for International Earth Science Information Network (CIESIN), Columbia University. http://www.africasoils.net/data/datasets (Accessed 13/02/2013).

Aitchison, J. 1990. Relative variation diagrams for describing patterns of compositional variability. *Mathematical Geology*, 22, 487–511.

Akamigbo F.O.R & Asadu, C.L.A. 1983. Influence of parent material on the soils of southeastern Nigeria. *East Africa Agricultural and Forestry Journal*, 48, 81-91.

Amhakhian, S.O & Achimugu, S 2011. Characteristics of soil on toposequence in Egume, Dekina Local Government Area of Kogi State. *Production Agriculture and Technology*, 7, 29-36.

Amusan, A. A., Olayinka, A., & Oyedele, D. J. 2005. Genesis, classification, and management requirements of soils formed in windblown material in the Guinea Savanna area of Nigeria. *Communications in soil science and plant analysis,* 36(15-16), 2015-2031.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J. Mendonça Santos, M.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., & Zhang, G. L. 2014. Chapter Three-GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. *Advances in Agronomy*, *125*, 93-134.

Ayuba, S.A., Akamigbo, F.O.R. & Itsegha, S.A. 2007. Properties of soils in River Katsina-

Ala catchments area, Benue State, Nigeria. *Nigerian Journal of Soil Science* 17(1):24-29

Bationo, A., Kihara, J., Vanlauwe, B., Waswa, B., & Kimetu, J. 2007. Soil organic carbon dynamics, functions and management in West African agro-ecosystems. *Agricultural systems*, *94*(1), 13-25.

Bennett, J.G.1980. Aeolian Deposition and soil parent materials in northern Nigeria. *Geoderma*: 24: 241-255.

Bishop, T.F.A., A.B. McBratney and G.M Laslett. 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, 91, 27–45, http://dx.doi.org/10.1016/S0016-7061(99)00003-8.

Bishop, T. F. A., McBratney, A. B., & Laslett, G. M. 2006. Digital soil-terrain modeling: the predictive potential and uncertainty. In: Grunwald, S. (Ed.) Environmental Soil-Landscape Modeling, CRC Press, Boca Raton, FL, USA. p. 185–213.

Bockheim, J.G. & Hartemink, A.E. 2013. Distribution and classification of soils with taxonomically defined clay-enriched horizons in the USA: a Review. *Geoderma*, 209, 153-160.

Breiman, L. 2001a. Random forests. *Machine Learning*, 45(1), 5–32.

Breiman, L. 2001b. Statistical modeling: The two cultures. *Statistical Science*, 16, 199-215.

Buchanan, S., Triantafilis, J., Odeh, I., & Subansinghe, R. 2012. Digital soil mapping of compositional particle-size fractions using proximal and remotely sensed ancillary data. *Geophysics*, 77(4), 201-211.

Chartres, C.J. 1982. The use of landform-soil associations in irrigation soil surveys in northern Nigeria. *Journal of Soil Science*, 33, 317-328.

Cook, S. E., Corner, R. J., Groves, P. R., & Grealish, G. J. 1996. Use of airborne gamma radiometric data for soil mapping. *Soil Research*, *34*(1), 183-194.

Cutler, D. R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. &. Lawler, J.J. 2007. Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

De Gruijter, J. J., Walvoort, D. J. J., & Van Gams, P. F. M. 1997. Continuous soil maps—a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma*, *77*(2), 169-195.

Dikau, R., Brabb, E.E. & Mark, R.M. 1991. Landform classification of New Mexico by computer. Open File Rep. 91-634. USGS, Denver, CO.

Diaz-Uriarte, R. & de Andres, S.A. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 3.

Federal Department of Agricultural Land Resources (FDALR) 1990. The reconnaissance soil survey of Nigeria. Volumes I-IV. FDALR, Nigeria.

Food and Agriculture Organization (FAO) 1984. Improved production system as an alternative to shifting cultivation. *Soils Bulletin* 53, FAO, Rome, Italy.

Gessler, P.E., McKenzie, N.J. & Hutchison, M.F. 1996. Progress in soil-landscape modelling and spatial prediction of soil attributes for environmental model. In: Proceedings, Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Santa Barbara, CA, National Centre for Geographic Information and Analysis. January 21-26.

Greve, M.H., Bou Kheir, R.., Greve, M.B. & Bøcher, P.K. 2012a. Using digital elevation models as an environmental predictor for soil clay contents. *Soil Science. Society of American Journal*, 76, 2116–2127. doi:10.2136/sssaj2010.0354

Greve, M.H., Bou Kheir, R.., Greve, M.B. & Bøcher, P.K. 2012b. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. *Ecological Indicators*, 18, 1–10. doi:10.1016/j.ecolind.2011.10.006.

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using Random Forests analysis. *Geoderma*, 146 (1–2), 102–113.

Grove, A.T. 1958. The ancient erg of Hausaland and similar formations on the south side of the Sahara. *Geographical Journal*, 124:528-533.

Hartemink, A.E. & McBratney, A.B. 2008. A soil science renaissance. *Geoderma*, 148, 123-129.

Hassan, A.M. 2010. Genesis, classification and agricultural potential of the soils derived from *Kerrikerri* sandstone formation in northern Nigeria. *Continental Journal of Agricultural Science*, 4, 7-19.

Hassink, J. 1997. The capacity of soils to preserve organic C and N by their association with clay and silt particles. *Plant and soil 191*(1), 77-87.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. The elements of statistical learning. Data Mining, Inference, and Prediction. Second edition. New York: Springer. 745p

Henderson, B. L., Bui, E. N., Moran, C. J., & Simon, D. A. P. 2005. Australia-wide predictions of soil properties using decision trees. *Geoderma*, *124*(3), 383-398.

Hekstra, P. & Andriesse, W. 1983. Wetland utilization research project, West Africa. Phase 1. The inventory. Volume II: Physical aspects. International Land Reclamation and Improvemnet (ILRI), Wageningen, The Netherlands. 78p

Higgins, G.M. 1963. Upland soils of Samaru and Kano plains. (Unnumbered Soil Survey. Report.). Institute for Agricultural .Research, Samaru, Nigeria.

Igwe C.A. 2005. Soil physical properties under different management systems and organic matter effects on soil moisture along soil catena in southeastern Nigeria. *Tropical. Subtropical. Agroecosystems*, 5, 57-66.

Igwe, C. A., Zarei, M. & Stahr, K. 2009. Mineralogy and geochemical properties of some upland soils from different sedimentary formations in south-eastern Nigeria. *Soil Research*, 47 (4), 423-432.

Iloeje, N.P. 2001. A new geography of Nigeria. New Revised edition. Longman Nigeria Plc. Lagos, Nigeria. 200p

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

Jalabert, S. S. M., Martin, M. P., Renaud, J. P., Boulonne, L., Jolivet, C., Montanarella, L., & Arrouays, D. 2010. Estimating forest soil bulk density using boosted regression modelling. *Soil Use and Management*, *26*(4), 516-528.

Jones, M.J., & Wild, A. 1975. Soils of the West African Savanna: The maintenance and improvement of their fertility. Commonwealth Agricultural Bureau, Farnham Royal, UK. 218p.

Kempen, B., Brus, D. & Stoorvogel, J.J. 2011. Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. *Geoderma,* 162, 107–123 doi:10.1016/j.geoderma.2011.01.010.

Klinkenberg, K. & Higgins, G.M. 1968. An outline of Northern Nigerian soils. *Nigerian Journal of Science*, 2(2), 91-115.

Kparmwang, T. 1993. Characterization and classification of basaltic soils in the Northern Guinea Savanna zone of Nigeria. Unpublished Ph.D thesis. Ahmadu Bello University, Zaria. 176p.

Law-Ogbomo, K.E. & Nwachokor, M.A. 2010. Variability in selected soil physic-chemical properties of five soils formed on different parent materials in southeastern Nigeria. *Research Journal of Agriculture and Biological Sciences*, 6(1), 14-19.

Lawes, D.A. 1962. The influence of rainfall conservation on the fertility of the loess plain soil of northern Nigeria. *Nigerian Geography Journal*. 5.

Leenaars, J.G.B. 2012. Africa Soil Profiles Database, version 1.0.: A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). ISRIC Rep. 3. ISRIC-World Soil Information, Wageningen, The Netherlands.

Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R news volume 2(3): 18-21.

Ließ, M., Glaser, B., & Huwe, B. 2012. Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models.*Geoderma*, *170*, 70-79.

Lombin, G., & Esu, I. 1988. Characteristics and management problems of Vertisols in the Nigerian savannah. Management of Vertisols in Sub-Saharan Africa, Proceedings of a Conference held at ILCA, 31 Aug. to 4 Sep., 1987.Addis Ababa, Ethiopia.

Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, *154*(1), 138-152.

Maniyunda, L. M., Raji, B. A. & Gwari, M. G. 2013. Variability of Some Soil Physicochemical Properties on Lithosequence in Funtua, North-Western Nigeria. *International Journal of Science and Research*, 2(9), 174-180.

Mayr, T.R., Palmer, R.C., & Cooke, H.J. 2008. Digital soil mapping using legacy data in the Eden valley, UK. In: Hartemink, A.E., McBratney, A.B., Mendonça Santos, M.d.L. (Eds.), Digital Soil Mapping with Limited Data. Springer Science+Business Media, pp. 291–301.

McBratney, A. B., Odeh, I. O., Bishop, T. F., Dunbar, M. S., & Shatar, T. M. 2000. An overview of pedometric techniques for use in soil survey.*Geoderma*, *97*(3), 293-327.

McBratney, A. B., Santos, M. M., & Minasny, B. 2003. On digital soil mapping. *Geoderma*, *117*(1), 3-52.

McTainsh, G.H. 1984. The nature and origin of the Aeolian mantles of central northern Nigeria. *Geoderma*, 33, 13-37.

Minasny, B., McBratney, A. B., Mendonça-Santos, M. L., Odeh, I. O. A., & Guyon, B. 2006. Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Soil Research*, *44*(3), 233-244.

Minasny, B., McBratney, A.B., & Lark, R. M. 2008. Digital soil mapping technologies for countries with sparse data infrastructures. In: Hartemink, A.E., McBratney, A.B., Mendonça Santos, M.d.L. (Eds.), Digital Soil Mapping with Limited Data. Springer Science+Business Media, p 15-30.

Minasny, B., & Hartemink, A.E. 2011. Predicting soil properties in the tropics. *Earth Science Revolution*, 106, 52–62. doi:10.1016/j.earscirev.2011.01.005.

Møberg, J., & Esu, I. 1991. Characteristics and composition of some savannah soils in Nigeria. *Geoderma*, 48, 113–129. doi:10.1016/0016-7061(91)90011-H.

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

Niang, M. A., Nolin, M. C., Jégo, G., & Perron, I. 2014. Digital Mapping of Soil Texture Using RADARSAT-2 Polarimetric Synthetic Aperture Radar Data. *Soil Science Society of America Journal*, *78*(2), 673-684.

Odeh, I. O. A., McBratney, A.B. & Chittleborough, D.J. 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, 67, 215-226.

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

Odeh, I. O. A., Todd, A. J. & Triantafilis, J. 2003. Spatial prediction of soil particle-size fractions as compositional data: *Soil Science*, 168, 501–515.

Odeh, I.O.A., Crawford, M & McBratney, A.B. 2007. Digital mapping of soil attributes for regional and catchment modelling, using ancillary covariates, statistical and geostatistical techniques. In: Lagacherie, P., A.B. McBratney & M.Voltz (eds). Digital soil mapping: An introductory perspective. *Development in Soil Science*, 31, 437-453.

Odeh, I.O.A., Leenaars, J.G.B., Hartemink, A. & Amapu, I. 2012. The challenges of collating legacy data for digital mapping of Nigerian soils. In: Minasny,B., Malone, B.P., McBratney,A. (eds). Digital Soil Assessments and Beyond. Taylor & Francis Group, London, p 453-458.

Ogban, P.I. & Babalola, O. 2003. Soil characteristics and contsraints to crop production in inland valley bottoms in southwestern Nigeria. *Agricultural Water Management*, 61, 13-28.

Ogunwale, J. A., Ashaye, T.I., Odu, C.T.I. & Fayemi, A.A.A. 1975. Characterization of selected sandstone-derived soils in the ecological zones of Nigeria. *Geoderma* 13, 331-347.

Ojanuga, A.G. 1975. Morphological, physical and chemical characteristics of soils of Ife and Ondo areas. *Nigerian Journal of Science*, 9, 225-256.

Ojanuga, A.G. 2006. Agroecological zones of Nigeria manual. National Special Programme for Food Security and FAO, Abuja, Nigeria.124p.

Oku, E., Essoka, A. & Thomas, E. 2010. Variability in soil properties along an Udalf toposequence in the humid forest zone of Nigeria. *Kasetsart Journal.of Natural Science,* 44, 564-573.

Okun, O. & Priisalu, H. 2007. .Random forest for gene expression based cancer classification: overlooked issues. In: Martı´, J., J.M. Benedı´, A.M. Mendonc and J. Serrat, (Eds.). Pattern Recognition and Image Analysis: Third Iberian Conference, Lecture Notes in Computer Science, Springer Berlin Heidelberg, p 483-490.

Olowolafe, E.O. 2002. Soil parent materials and soil properties in two separate catchment areas on the Jos Plateau, Nigeria. *GeoJournal*, 56, 2001-212.

Omoregie, A.U.1998. Selected physico-chemical characteristics of soils from the sub-humid and semi-arid zones of Nigeria, *Communications in Soil Science and Plant Analysis*, 29, 15-16, 2505-2513, DOI: 10.1080/00103629809370128

Osei, B. A., & Okusami, T. A. 1994. Classification of soils derived from amphibolite parent material in south-western Nigeria. *Ghana Journal of Agricultural Science,* 24(1), 123-132.

Oyedele, D. J., Schjønning, P & Amusan, A. A. 2006. Physicochemical properties of earthworm casts and uningested parent soil from selected sites in southwestern Nigeria. *Ecological Engineering,* 28(2), 106-113.

Prasad, A. M., Iverson, L. R. & Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9, 181-199.

Priori, S., Bianconi, N. & Costantini, E.A. 2014. Can γ-radiometrics predict soil textural data and stoniness in different parent materials? A comparison of two machine-learning methods. *Geoderma*, *226*, 354-364.

R Development Core Team 2013. R: A language and environment for statistical computing, R version 2.15.3. R Foundation for Statistical Computing, Vienna, Austria.

Rawlins, B. G., Marchant, B. P., Smyth, D., Scheib, C., Lark, R. M., & Jordan, C. 2009. Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland. *European Journal of Soil Science*, *60*(1), 44-54.

Reuter, H.I., & Nelson, A. 2009. Geomorphometry in ESRI Packages. In: Hengl T. and H.I. Reuter (Eds), Geomorphometry: Concepts, software, applications, Dev. Soil Sci. 33. Elsevier, New York. p 269-291.

Salako, F. K., Tian, G., Kirchhof, G., & Akinbola, G. E. 2006. Soil particles in agricultural landscapes of a derived savanna in southwestern Nigeria and implications for selected soil properties. *Geoderma*, *137*(1), 90-99.

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.D., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., & Zhang, G.L., 2009. Digital soil map of the world. *Science*, 325, 680-681.

Scull, P., Okin, G., Chadwick, O. A. & Franklin, J. 2005. "A comparison of methods to predict soil surface texture in an alluvial basin. *The Professional Geographer*, 57(3), 423-437.

Sharu, M. B., Yakubu, M., Noma, S. S., & Tsafe, A. I. 2013. Land Evaluation of an Agricultural Landscape in Dingyadi District, Sokoto State, Nigeria.*Nigerian Journal of Basic and Applied Sciences*, *21*(2), 148-156.

Smyth, A.J. and R.F. Montgometry. 1962. Soils and land use in central western Nigeria. Government of Western Nigeria Press, Ibadan, Nigeria. 265p.

Soil Survey Staff. 2006. Keys to soil taxonomy. 10th edition. NRCS, Washington, DC.

Soil Survey Staff. 2010. Keys to Soil Taxonomy (eleventh ed.)USDA-Natural Resources Conservation Service, Washington, DC

Sombroek, W.G. & Zonneveld, I.S. 1971. Ancient dune fields and fluviatile deposits in the Rima-Sokoto river basin soil survey Paper. 5. Netherlands. Soil Survey Institute, Wageningen.

Tamminen, P. & Starr, M. 1994. Bulk density of forested mineral soils. *Silva Fennica*, 28, 53–60.

Thompson, J. A., Pena-Yewtukhiw, E. M. & Grove, J. H. 2006. "Soil–landscape modeling across a physiographic region: Topographic patterns and model transportability." *Geoderma*, 133(1), 57-70.

Thompson, J. A., Roecker, S., Grunwald, S. & Owens, P. R. 2012. Digital soil mapping: Interactions with and applications for Hydropedology. In: Lin, H. (Ed.). Hydopedology, First edition, Elsevier. p 665-709.

Tomlinson, P.R. 1961. Report on the detailed soil survey of the Livestock Investigation Centre, Katsina, and the reconnaissance survey of the surrounding area. Soil Surv. Publ. 11. Institute for Agricultural Research, Samaru, Nigeria.

Udo, R. K. 1970. Geographical regions of Nigeria. Heinemann, London. 212p.

Ugbaje, S.U. & Reuter, H.I. 2013. Functional digital soil mapping for the prediction of available water capacity in Nigeria using legacy data. *Vadoze Zone Journal*: doi:10.2136/vzj2013.07.0140.

van den Boogaart, K. G. & Tolosana-Delgado, R. 2008. "Compositions": a unified R package to analyze compositional data. *Computers & Geosciences*, 34(4), 320-338.

Vasques, G. M., Grunwald, S., Comerford, N. B., & Sickman, J. O. 2010. Regional modelling of soil carbon at multiple depths within a subtropical watershed. *Geoderma*, *156*(3), 326-336.

Vine, H. 1987. Wind-blown materials and W African Soils: an explanation of the 'ferrallitic soil over loose sandy sediments' profile. *Geological Society, London, Special Publications* 35, 171-183. Doi: 10.1144/GSL.SP.1987.035.01.12.

Chapter 4.

Total soil organic carbon and carbon sequestration potential in Nigeria

Published: **Akpa, S. I.C.,** I.O.A. Odeh, T.F.A. Bishop, A.E. Hartemink, & I.Y. Amapu 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma,* 271, 202-215.

Abstract

This study aimed to quantify SOC stocks and potential C sequestration for Nigeria using legacy soil data. Mass preserving splines were fitted to legacy SOC and bulk density (BD) pedon data based on GlobalSoilmap soil depths. SOC concentrations (g kg^{-1}) were predicted using Random Forest Model (RFM), *Cubist* and Boosted Regression Tree (BRT). Thereafter, the soil carbon density (Mg C ha^{-1}) was calculated from the SOC concentration and BD (Mg m^{-3}). The information was combined with land use/land cover (LULC) map and agro-ecological zone (AEZ) digital maps to estimate SOC sequestration. The mean SOC concentration ranged between 4.2 and 23.7 g kg^{-1} in the top 30 cm and between 2.6 and 9.2 g kg^{-1} at the lower soil depth. Total SOC stock in the top 1 m was 6.5 Pg with an average density of 71.60 Mg C ha^{-1}. Almost half of the SOC stock was found in the 0-30 cm layer. SOC stocks decreased from the southwest to the northeast of Nigeria, and increased from Sahel to Humid forest AEZs. Restoration of the various land use types has the potential to sequester about 0.2 to 30.8 Mg C ha^{-1} depending on the AEZ. The Derived Guinea Savannah presents a potential hotspot for targeted carbon sequestration projects in Nigeria. Knowledge of SOC stock and sequestration is vital for framing appropriate management regimes to increase soil carbon stocks and for C accounting purposes.

Key words

Digital soil mapping, Legacy data, Soil organic carbon, Random Forest, Cubist, Nigeria.

4.1 Introduction

There is a growing concern over the contribution of agricultural sector to the increasing global warming (IPCC, 2011). This concern has heightened demand for information on spatial patterns of soil organic carbon (SOC) stocks in relation to agricultural land uses and land use/cover (LULC) change. LULC change and the management of agro-ecosystems have the potential to release considerable amount of SOC stored in the soil through tillage, cropping systems, irrigation, fertilization and other agricultural operations (Bruce et al. 1999; Lal, 2005). The soil carbon pool constitutes about two-third of the total terrestrial carbon pool, which is three times the amount of atmospheric carbon (Smith, 2012). Thus it is important to decipher the spatial distribution of soil carbon stock to identify where anthropogenic factors are contributing significantly to carbon-dioxide (CO_2) emissions into the atmosphere

SOC is sensitive to changes in land use (Poeplau and Don, 2013) and a change from natural or semi-natural LULC to agricultural ecosystems often leads to significant changes in SOC content (Post and Kwon, 2000; Guo and Gifford, 2002; Wilson et al., 2008). According to Powers et al. (2011) the conversion of forests to shifting cultivation or permanent crops can reduce SOC stocks by an average of almost 20% over a period of time. Other studies have estimated the loss of SOC after cultivation of virgin land to be between 20% and 50% (Post & Kwon 2000; Guo & Gifford 2002; Murty et al. 2002; Gregorich et al. 2005). Overall, long-term agricultural land use change could decrease soil C content by 48% in the top 10 cm (Don et al., 2011; Poeplau et al., 2011) with a concomitant increase in atmospheric C.

In contrast to land clearing, land management can preserve the SOC pool or even lead to increased C sequestration and thus reduced atmospheric CO_2 concentration (Jenny 1980; Post et al., 1998; Metting et al., 1999). Additionally, increased carbon storage could be achieved through afforestation where low biomass LULC types such as grasslands or croplands are converted to forests and plantations (Roshetko et al., 2007; Nave et al., 2013). Besides sequestration of C in the soil through C input, afforestation causes increased stabilization of old C as fine fractions protected by micro-aggregates (Del Galdo et al., 2003; Mulugeta et al., 2005; Bekele et al., 2006). Other studies have demonstrated that the conversion of forest to well-managed pastures can enhance SOC storage compared to SOC storage under native forest (Powers et al. 2011). Increased SOC storage following the conversion of cropland to grassland has also been reported (Su et al., 2009; Fang et al., 2012; Poeplau and Don 2013).

The capacity of soils to store more C following restoration of various land uses to their pristine ecosystem, depends on several factors such as vegetation (Jobbagy and Jackson, 2000), climatic conditions (Dixon et al., 1994), soil texture (Six et al., 2002) and topography (Rosenbloom et al., 2006). Climatic elements affect SOC storage through alteration of decomposition rate of SOC as well as changes in the quantity and quality of C cycled through the ecosystem. Vegetation often determines the vertical distribution of SOC through root biomass differences with depth (Jobbagy and Jackson, 2000; Dorji et al., 2014). In addition to climate and vegetation, soil properties, such as texture, play important role in C storage through their stabilizing effects on SOC (Jobbagy and Jackson, 2000). Also, topography affects SOC stock through its influence on soil moisture regime as well as redistribution of soil particles (Gulledge and Schimel, 2000).

Although several studies have shown that LULC changes affect SOC content and sequestration of soils (Post and Kwon, 2000; Albaladejo, et al., 2013), the magnitude and dynamics of these changes in different ecosystems have not been extensively studied. In Nigeria, for example, natural ecosystems have been degraded following deforestation, overgrazing, nutrient mining, soil erosion, and loss of bio-diversity (UNEP, 2007). These degraded lands have great potential to sequester C in the soils (Follett et al., 2001). In addition, most soils in Nigeria are highly weathered with low activity clays (FMANR, 1990) that have small mineral surfaces to allow physical protection and stabilization of SOC. Such soils are more susceptible to perturbations associated with LULC changes, leading to SOC decline. Several studies on the influence of land use on SOC storage have been reported for various ecosystems in

Nigeria (Raji and Ogunwole, 2006; Anikwe, 2010; Obalum et al., 2012). These studies are localized based on small datasets and no information on SOC storage up to 1m soil depth has been covered. This presents uncertainties in the understanding of the impacts of LULC change on the C cycle and the sustainability of agricultural systems (Meersmans et al., 2009; Wiesmeier et al., 2012). This study therefore aims to (i) estimate the total SOC stock and (ii) determine the potential carbon sequestration of soils under different land use types across agro-ecological zones of Nigeria.

4.2 Materials and methods

4.2.1 Study Area

Nigeria, with a total area of about 923,768 km^2, extends across a broad geographical area characterised by a large climatic range with two major biomes: the tropical humid forest in the south, and the savannah in the north (Keay, 1959). The savannah comprises Southern Guinea, Northern Guinea, Sudan, and Sahel zones respectively (Adegbehin and Igboanugo, 1990**)**. An addition to the two vegetation types is the derived savannah which is a transition zone between the rainforest and savannah caused by significant loss of forest by clearance. These climatic and vegetative variations, combined with the soil, constitute the agro-ecological zones (AEZs) shown in Table 4.1 (IITA, 1992). The environmental and anthropogenic factors across these AEZs give rise to a somewhat north-south gradient in LULC across Nigeria (see Fig. 4.1). LULC ranges from sparse vegetation and grassland in the fringes of the northern region, through cropland/savannah/shrubland mosaics in the middle belt region to cropland/shrubland/forest mosaics in the coastal southern region. The LULC distribution includes cropland (31 %), Savanna (36%), grassland (18%), forest (11%), shrubland (1%) and others (3%).

Farming systems in Nigeria are heterogeneous depending on the agro-ecological and socio-economic environments. This is exacerbated by the variability in farmers' land holdings and farm management (Giller et al., 2011). Farming systems range from shifting cultivation and perennial tree cropping in the humid forest AEZ to crop–livestock farming in the savannah (Dixon et al., 2001). Smallholder farming systems are variable with farm size ranging from 0.2 to less than 2 ha.

Table 4.1 Description of the major agro-ecological zones (AEZ) in Nigeria.

Agro-ecological Zones	Annual rainfall (mm)	Annual Temperature (^oC)	Days of growing period	Pristine vegetation	Dominant soils (WRB)
Humid Forest	2000-3000	25-27	270-360	Forest	Ferralsols, Acrisols
Derived Guinea Savannah	1500-2000	26-28	211-270	Forest	Ferralsols, Luvisols, Arenosols, Nitosols
Southern Guinea Savannah	1200-1500	26-29	181-210	Savannah	Luvisol, Ferralsols, Acrisols, Lithosols
Northern Guinea Savannah	900-1200	27-29	151-180	Savannah	Luvisols, Vertisols, Lithosols,Ferralsols
Sudan Savannah	500-900	25-30	91-150	Savannah	Lixisols, Luvisols, Regosols
Sahel Savannah	250-500	21-32	≤90	Grassland	Aridisols, Regosols
Montane/High Altitude	1100-1500	20-23	160-200	Savannah	Luvisols, Lithosols, Ferralsols

†Adapted and Modified from Sowunmi & Akintola. (2010) and Jagtap, 1995.

4.2.2 Data sources and processing

4.2.2.1 Soil data

The major SOC and BD profile data used in this study were taken from the ISRIC compilation of Africa Soil Profiles Database obtained from soil survey reports and field research conducted in Nigeria (Leenaars, 2012; Odeh et al., 2012). The procedures for the determination of these properties were already described by Leenaars (2012). Bulk of the SOC contents data were measured by wet oxidation/digestion using either the Walkley Black (WB) method or the modified WB method of Nelson and Sommers (1996), while a few were measured using the dry combustion method. However, data obtained from these two methods were harmonized to the dry combustion method (Leenaars, 2012). The BD data were determined using the core sampling method. The SOC concentrations were originally reported in percentage mass unit but were converted to g kg^{-1} following the GlobalSoilMap specifications (Arrouays et al., 2014). We also calculated mass-preserving splines (Bishop et al., 1999) to convert the soil profile data to standard depth intervals (0-5, 5-15, 15-30, 30-60, 60-100, 100-200 cm) in accordance with the GlobalSoilMap specifications (Arrouays et al., 2014). Overall, SOC data from 711 soil profiles and BD data from 222 profiles were used in this study after data pre-processing. The distribution of the SOC profile data across the various AEZs in Nigeria is shown in Fig. 4.2.

4.2.2.2 Predictor variables

In this study, 23 predictor variables were used in the SCORPAN model (McBratney et al., 2003) as the predictors of SOC and BD- both of which are fundamental to SOC stock estimation (see Table 4.2). The predictors include SRTM 90 m digital elevation model (DEM) (USGS, 2006) from which other predictors, such as slope gradient, aspect, profile and plan curvatures, flow accumulation, topographic wetness index (TWI), stream power index (SPI), were derived following Reuter and Nelson. (2009). Related predictors used include landform classifications based on algorithms by Iwahashi (Iwahashi and Pike, 2007) and Hammond (Dikau et al., 1991), physiographic regions map derived from DEM (Akpa et al 2014), MODIS enhanced vegetation index (EVI) and Normalized difference Vegetation Index (NDVI) maps (obtained from https://lpdaac.usgs.gov), and bands 1, 2, 3, 4 and 7 of Landsat 7-ETM+ coverage obtained from Landsat GeoCover ETM+ 2000 edition (MDA Federal, 2004). To complete the picture, SCORPAN predictor variables of mean annual rainfall and temperature data acquired from the 1km global climate data (Hijmans et al., 2005) soil map of Nigeria (FMANR, 1990) and generalized geology map of Nigeria digitized from the UNESCO geology map of Africa (UNESCO-ASGA, 1963) were included. All these data layers were first transformed to a common projection (UTM WGS84 Zone 32N) and then resampled to 1000m resolution using the nearest neighbour technique in ArcGIS10.1 prior to further analysis.

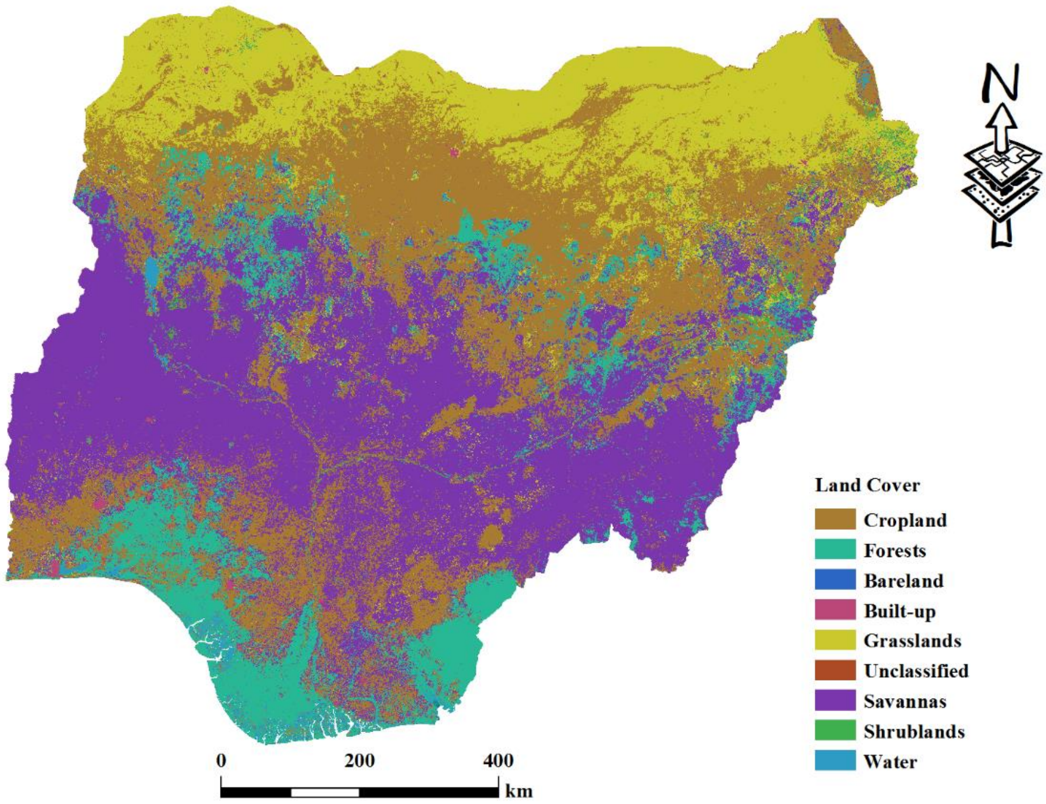


Figure 4.1. Generalized Land use/land cover map of Nigeria Reclassified from MODIS Global land cover (Friedl et al., 2010).

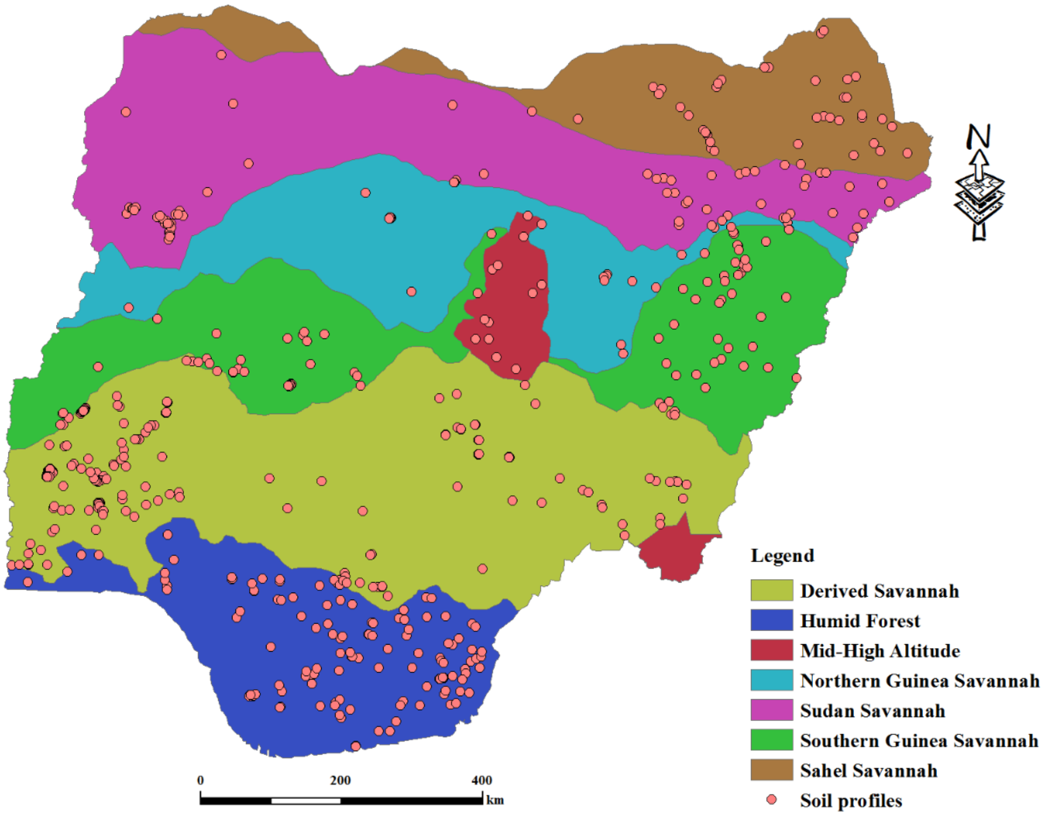


Figure 4.2. SOC legacy profile distribution across the various Agro-ecological zones in Nigeria (AEZ map adapted from IITA, 1992 and Jagtap, 1994).

4.2.3 Modelling and spatial prediction

4.2.3.1 Prediction models

We employed three non-parametric prediction models (Random forest, *Cubist* and Boosted regression tree) for the spatial prediction of SOC and BD. Prior to fitting each of the three prediction models, the “train” function of the “caret” R package was used to obtain optimal parameter settings for each of the models at each depth interval. The train function has the capacity to fine tune various models by selecting a combination of sensitive parameters that are associated with the optimal resampling statistics of the held-out samples (Akpa et al., 2014).

Table 4.2. Predictor variables used to predict SOC and Bulk density for Nigeria.

Variables	Data source	Original Scale/Resolution	References
Topography			
Slope Aspect	SRTM DEM	90 m	USGS (2006)
Slope gradient	SRTM DEM	90 m	USGS (2006)
Elevation	SRTM DEM	90 m	USGS (2006)
Wetness index	SRTM DEM	90 m	USGS (2006)
Stream power index	SRTM DEM	90 m	USGS (2006)
Flow Accumulation	SRTM DEM	90 m	USGS (2006)
Plan curvature	SRTM DEM	90 m	USGS (2006)
Profile curvature	SRTM DEM	90 m	USGS (2006)
Physiographic region	SRTM DEM	90 m	Akpa et al. (2014)
Iwahashi	SRTM DEM	90 m	Akpa et al. (2014)
Hammond	SRTM DEM	90 m	Akpa et al. (2014)
Vegetation/Anthropogenic factors			
Landuse	MODIS	500 m	https://lpdaac.usgs.gov
NDVI	MODIS	250 m	https://lpdaac.usgs.gov
EVI	MODIS	250 m	https://lpdaac.usgs.gov
Band 1	Landsat	30 m	MDA Federal, 2004
Band 2	Landsat	30 m	MDA Federal, 2004
Band 3	Landsat	30 m	MDA Federal, 2004
Band 4	Landsat	30 m	MDA Federal, 2004
Band 7	Landsat	30 m	MDA Federal, 2004
Climate			
Precipitation	WorldClim data	1 km	Hijmans et al. (2005)
Temperature	WorldClim data	1 km	Hijmans et al. (2005)
Parent Material			
Geology	Scanned and digitized geological map	1:5,000,000	UNESCO-ASGA, 1963
Soil types	FMANR	1:650,000	FMANR, 1990

†FMANR; Federal department of agriculture and land resources, DEM; digital elevation model, NDVI; normalized difference vegetation index, EVI; enhanced vegetation index, MODIS; moderate resolution imaging spectroradiometer, SRTM; shuttle radar topography mission

4.2.3.1.1 Random Forest model

Random forest model (RFM0 is a tree-based, robust prediction technique which was developed by Breiman (2001) but only recently employed in digital soil mapping (DSM) studies (Grimm et al., 2008). RFM has been successfully used in spatial prediction of SOC stocks because the underlying tree models can accommodate non-linearity in the response-predictor relationship, and interactions between the predictors (Grimm et al., 2008; Wiesmeier et al., 2011; Vågen et al., 2013). The model's strength lies in its two randomization procedures of bootrapping and random input selection. In addition, RFM carries out bagging of predictions which subsequently improves predictions of the individual tree models (Suuster et al., 2012; Vaysse and Lagacherie, 2015). We implemented RFM in spatial prediction of both SOC and BD using the *randomForest* 4.6 package in R environment (R Development Core Team, 2014).

4.2.3.1.2 Cubist

Cubist is a rule based model that is an extension of Quinlan's M5 model tree (Quinlan, 1993). The approach used in *Cubist* for tree growing is similar to those used in classical regression tree models such as classification and regression trees (CART). However, unlike CART the terminal tree leaves contain linear regression models instead of discrete class labels (Minasny and McBratney, 2008). In *Cubist*, regression trees are further reduced to a set of comprehensible rules, with each rule based on some conditions so that different linear models are able to capture local linearity in the predictor variable space, thus leading to smaller trees and better prediction accuracy when compared with CART (Quinlan, 1993). *Cubist* has gained wide application in DSM recently especially in SOC stock modelling (Bui et al., 2009; Miklos et al., 2010; Stevens et al., 2013; Adhikari et al., 2014; Lacoste et al., 2014; Mulder et al., 2016). We carried out *Cubist* modelling of SOC and BD using the *Cubist* package in R environment (Kuhn et al., 2013).

4.2.3.1.3 Boosted Regression Tree

Boosted regression trees (BRT) belong to the gradient boosting modelling family of statistical algorithms (Collard et al., 2014). Like CART, it builds regression trees to make prediction of a target variable but improve prediction accuracy by minimizing the risk of over-fitting through boosting technique (Lawrence et al., 2004). Boosting techniques are generally applied to increase performance of a given estimation method by generating instances of the method iteratively from a training data set and additively combining them in a forward “stage-wise” procedure (Elith et al., 2008). Like most data mining prediction models BRT has an inherent ability to represent interactions among predictor variables without a priori knowledge of their distribution. Additionally, BRT is robust to the effects of outliers, missing data and autocorrelation among variables (Jalabert et al., 2010). Two main parameters are required for the fitting of BRT: the learning rate and the tree size or interaction depth. We applied BRT in modelling SOC and BD using the “gbm” package in the R statistical environment.

4.2.3.2 Model evaluation

The performance of the aforementioned three models in predicting SOC and BD was tested by cross-validation, with 80% of the data used for model calibration while the remaining 20% for model validation. To ensure stability and increase reliability, model calibration was based on 100 iterations or runs. Each run involved random sampling of the subsets for calibration and validation after which the performance of each model was evaluated using the difference between measured and predicted response variable. In doing this, three statistical indices: root mean square error (RMSE), coefficient of determination (R^2) and Lin’s concordant correlation coefficient (Pc), were computed.

RMSE, which is a measure of model accuracy, was computedas:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \qquad [4.1]$$

where, n denotes data points, o_i and p_i are observed and predicted SOC concentration and BD values at the ith point.

The coefficient of determination (R^2), which is the percentage of variation explained by each model, was calculated as:

$$R^2 = \frac{\sum_{i=1}^{n}(p_i - \mu_o)^2}{\sum_{i=1}^{n}(o_i - \mu_o)^2} \qquad [4.2]$$

Where, μ_o and μ_p are the means for the raw and predicted SOC and BD as the case may be.

The Lin's concordance correlation coefficient (ρ_c), a measure of the strength of the agreement between the observed and predicted PSF values, was computed as:

$$\rho_c = \frac{2\rho\sigma_o\sigma_p}{\sigma_o^2 + \sigma_p^2 + (\mu_o - \mu_p)^2} \qquad [4.3]$$

where, ρ_c is the estimated Lin's concordance correlation coefficient, μ_o and μ_p are the means for the raw and predicted SOC and BD while σ_o^2 and σ_p^2 are the corresponding variance and ρ the Pearson correlation coefficient between the raw and predicted SOC and BD. A good model will have a ρ_c close to 1 and RMSE of almost 0.

4.2.3.3 Prediction uncertainty

One of the strengths of DSM is the quantification of the uncertainty inherent in spatial prediction. Since models are representations of reality (Luoto and Hjort, 2005), prediction of soil attributes have a level of uncertainty that can be quantified. This is important for guiding decision-making processes (Goovaerts, 2001). We therefore estimated the uncertainty of our predictions by calculating 95% prediction intervals from the individual bootstrap predictions of the numerous trees or rules generated by the different models for both SOC and BD at each of the five depth intervals. The 95% prediction intervals for SOC and BD were then used to calculate uncertainty in our SOC stock estimation.

4.2.4 Estimation of SOC density and stocks

SOC density (SOCD), which is the SOC mass per unit area for a given depth, was estimated for each depth interval as the product of SOC concentration, thickness of the layer interval and the bulk density using equation (4.4) below:

$$SOCD = SOC_c * BD * D * 10 \qquad [4.4]$$

where, $SOCD$ is SOC density (Mg ha^{-1}), SOC_c is SOC concentration (g kg^{-1}) in oven-dry basis, BD is bulk density (Mg m^{-3}) and D is depth interval thickness (m).

SOC stock (SOCS) is the actual SOC mass for a given soil depth and area. It was calculated by summing up the product of SOC density and area of the grid cell size as:

$$SOCS = \sum_{i=1}^{n} \left\{ \left(SOCD_i * A_i \right) / 10^6 \right\} \qquad [4.5]$$

where, $SOCS$ is the SOC stock (Tg), n is the number of grid cells, $SOCD_i$ is SOC density (Mg ha^{-1}) per grid cell for a given depth interval, A_i is the area of grid cell (ha) and 10^6 is the conversion factor from Mg to Tg.

4.2.5 Potential carbon sequestration

To estimate potential carbon sequestration, we calculated the difference between the mean total SOC of soils in the different LULC types and the respective pristine vegetation of a given AEZ. The assumption here is that, this will indicate the amount of C that could be sequestered when any of the other land use types is converted back to the pristine vegetation. This is can be represented as:

$$C_{seq} = \sum_{j=1}^{n} \left(\bar{SOCDn}_j - \bar{SOCDu}_j \right) \quad [4.6]$$

Where, C_{seq} is the potential C sequestration (Mg C ha^{-1}), $\bar{SOCDn}_j$ is mean SOCD (Mg C ha^{-1}) for native vegetation in a given AEZ and $\bar{SOCDu}_j$ is the mean SOCD (Mg C ha^{-1}) for any other land use type within the same AEZ.

4.3 Results

4.3.1 Modelling of SOC concentration and bulk density

4.3.1.1 Performance of prediction models

The performances of RFM, Cubist and BRT in predicting soil organic carbon and bulk density based on average values of 100 model runs cross-validation are shown in Table 4.3. In terms of R^2 and ρ_c, RFM and Cubist model exhibited similar performance although each out-performed BRT in predicting SOC especially the topsoil SOC. However, while RFM performed slightly better than Cubist model in terms of prediction error, Cubist performance is slightly better than RFM in terms of R^2 and ρ_c in predicting BD especially of the subsoil. The RFM captured 18 to 34% of the variation in SOC concentration and 25 to 48% of the variation in bulk density. This was higher in the top 15cm soil depth for SOC than the lower depth intervals. The RMSE ranged from 2.29 g C kg^{-1} to 7.96 g C kg^{-1} for SOC and 0.13 Mg m^{-3} to 0.16 Mg m^{-3} for the BD. Lin's concordance coefficients (Pc) range from 0.30 to 0.47 for SOC and from 0.35 to 0.57 for BD.

Each of the three models showed that soil type, climate, vegetation indices and terrain attributes are important predictors of SOC (see Fig 4.3) while soil type, climate, terrain attributes and soil surface reflectance indices are important predictors of BD in this study (see Fig. 4.4).

4.3.1.2 SOC concentration and bulk density

The summary of the spline-fitted and predicted SOC and BD for the three models is presented in Table 4. SOC contents are relatively low with mean ranging from 3.2 to 10.9 g C kg^{-1} (Spline fitted values) and are highly variable within the profile (SD of 2.5 to 9.7 g C kg^{-1}) especially for the 0-30 cm soil depth. SOC ranged from 0.6 to 102.7 g C kg^{-1} in the topsoil (0-30 cm) and from 0.1 to 42.8 g C kg^{-1} in the subsoil. Generally mean SOC contents decreased with depth. The predicted SOC by the three models show similar trend as the spline fitted data. However, the mean values of SOC predicted by RFM and *Cubist* are closer to the spline fitted data than the SOC predicted by BRT, with the latter over predicting values.

Table 4.3. Performance models in predicting soil organic carbon and bulk density.

Depth (cm)	RFM			Cubist			BRT		
	RMSE	R^2	ρ_c	RMSE	R^2	ρ_c	RMSE	R^2	ρ_c
Soil organic carbon (g kg^{-1})									
0-5	7.96	0.34	0.47	8.24	0.32	0.48	8.52	0.25	0.32
5-15	6.71	0.30	0.43	7.32	0.26	0.42	7.39	0.20	0.29
15-30	5.57	0.20	0.30	5.90	0.16	0.26	5.57	0.11	0.20
30-60	3.20	0.20	0.30	3.45	0.13	0.22	3.46	0.09	0.15
60-100	2.29	0.18	0.30	2.40	0.11	0.21	2.34	0.07	0.13
Bulk density (Mg m^{-3})									
0-5	0.16	0.25	0.35	0.16	0.20	0.36	0.17	0.17	0.17
5-15	0.14	0.29	0.39	0.15	0.24	0.40	0.15	0.21	0.21
15-30	0.13	0.36	0.46	0.14	0.32	0.49	0.14	0.29	0.28
30-60	0.13	0.44	0.54	0.13	0.43	0.60	0.15	0.35	0.36
60-100	0.14	0.48	0.57	0.14	0.48	0.65	0.15	0.42	0.44

† BRT, Boosted regression trees, ρ_c; Lin's concordance correlation coefficient, RFM; Random forest model, RMSE; Root mean square error

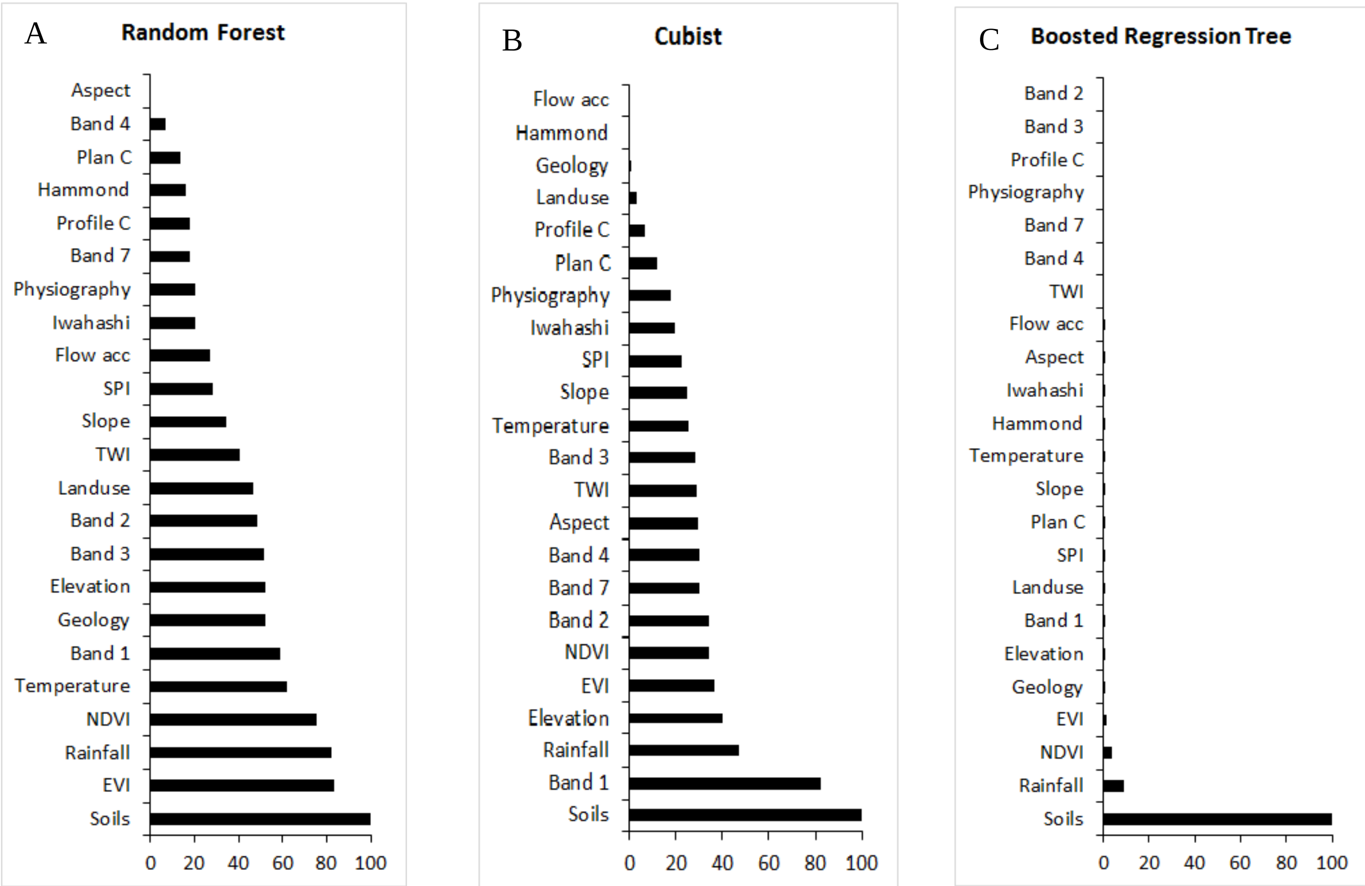


Figure 4.3. Importance of predictor variables in predicting SOC at the top 15 cm soil depth based on Random forest model (A), Cubist (B) and Boosted regression tree (C). Abbreviations: EVI; enhanced vegetation index, SPI; stream power index, NDVI; normalized difference vegetation index, TWI; topographic wetness index, Profile C; profile curvature, Plan C; plan curvature.

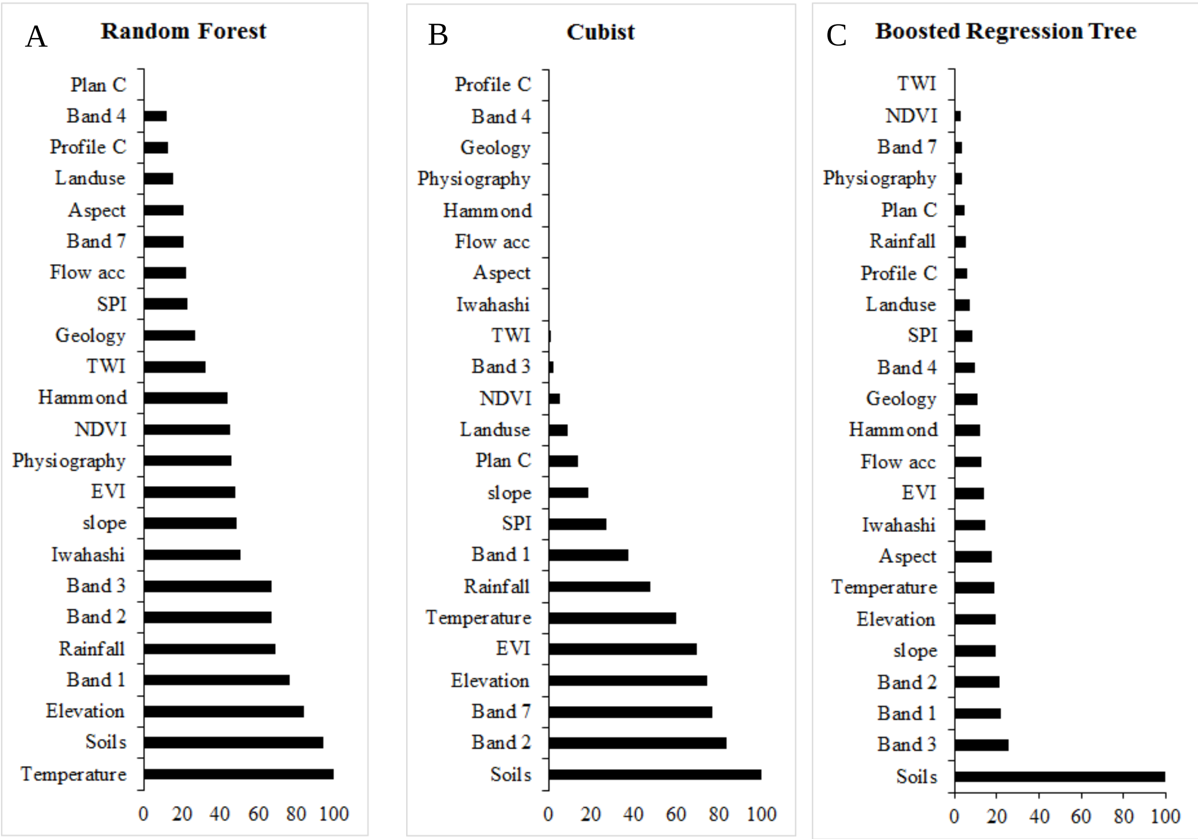


Figure 4.4. Importance of predictor variables in predicting bulk density at the top 15 cm soil depth based on Random forest (A), Cubist (B) and Boosted regression tree (C). Abbreviations: EVI; enhanced vegetation index, SPI; stream power index, NDVI; normalized difference vegetation index, TWI; topographic wetness index, Profile C; profile curvature, Plan C; plan curvature

Table 4.4. Summary statistics of spline-fitted and predicted soil organic carbon (g kg-1) and bulk density (Mg m-3) based on Random forest, Cubist and Boosted regression trees.

	RFM				Cubist model				BRT				Spline			
Depth (cm)	max	min	mean	SD	max	min	mean	SD	max	min	mean	SD	max	min	mean	SD
							Soil organic carbon									
0-5	47.0	3.40	11.0	5.30	70.6	0.62	10.5	6.50	47.1	13.3	18.3	2.50	93.3	0.70	10.9	9.70
5-15	44.4	3.10	9.80	4.30	81.2	0.44	9.40	5.70	45.2	8.50	17.7	2.60	102.7	1.00	9.60	8.30
15-30	31.8	2.40	7.40	2.70	115.8	0.43	6.90	3.90	29.0	8.00	14.3	0.90	94.6	0.60	7.00	6.10
30-60	18.9	1.80	4.80	1.50	50.6	0.82	4.60	2.00	17.0	5.50	9.20	0.60	42.8	0.30	4.50	3.60
60-100	11.2	1.20	3.40	0.90	25.2	-0.66	3.30	1.70	13.1	5.20	8.30	0.90	26.1	0.20	3.20	2.50
							Bulk density									
0-5	1.55	1.06	1.31	0.07	2.95	0.64	1.35	0.17	1.45	1.22	1.29	0.02	1.84	0.73	1.30	0.18
5-15	1.55	1.08	1.32	0.07	2.55	0.72	1.34	0.16	1.37	1.17	1.22	0.01	1.83	0.74	1.31	0.17
15-30	1.52	1.11	1.33	0.07	2.30	0.47	1.35	0.16	1.40	1.17	1.22	0.02	1.80	0.78	1.32	0.16
30-60	1.57	1.11	1.33	0.08	2.38	0.63	1.36	0.16	1.34	1.16	1.22	0.01	1.84	0.86	1.33	0.17
60-100	1.58	1.03	1.31	0.10	2.33	0.51	1.34	0.15	1.47	1.17	1.29	0.04	1.81	0.87	1.31	0.19

†BRT; Boosted regression trees, Max; maximum, Min; Minimum, SD; standard deviation, RFM; Random forest model.

The BD shows a low variability (SD range of 0.16 to 0.19 Mg m^{-3}) within the profile (spline fitted values). BD values ranged from 0.73 to 1.84 Mg m^{-3} in the surface layers and from 0.86 to 1.84 Mg m^{-3} in the subsurface layers. The mean BD is relatively uniform, ranging from 1.31 to 1.33 g Mg m^{-3} across all soil depth intervals. Among the three models RFM predicted more similar values of BD with the raw data compared to *Cubist* and BRT. *Cubist* over-predicted BD while BRT under-predicted BD.

4.3.2 SOC concentration of land use types

The vertical distribution of SOC content under different land use types based on RFM and *Cubist* is presented in Fig. 4.5. The models predicted SOC similar to the spline-fitted data except for model smoothing effect. Mean SOC predicted by both models is slightly lower in soils under forest (FL) and shrubland (SL) than spline-fitted SOC data especially at the top 30 cm depth while *Cubist* tends to under-predict SOC under grassland especially at the topsoil. The trend of SOC distribution across the various LULC types for bot model follows the order forestland > cropland > shrubland > savanna > grassland in the top 15cm depth. However, Shrubland and savanna have a higher SOC concentration compared to cropland below 30 cm soil depth. Overall, SOC content ranges from 1.8-47.0 g C kg^{-1} (FL), 2.0-31.3 g kg^{-1} (SL), 1.3-44.4 g kg^{-1} (S), 1.5-33.4 g kg^{-1} (GL), and 1.7-44.2 g kg^{-1} (CL).

4.3.3 SOC density and stock for Nigeria

The spatial distribution patterns of the total SOC density (SOCD) based on RFM and *Cubist* are shown in Fig. 4.6. RFM and Cubist yielded a similar trend in spatial distribution of SOCD. Generally, SOCD vary greatly across the study area; decreasing from the southwest to the northeast, and increasing from Sahel Savannah to Humid Forest agro-ecological zones. An average density of 71.60 Mg C ha^{-1}

accumulated in the top 1 m of the soils in Nigeria. Expectedly, SOCS shows spatial distribution as SOCD across the entire country. About 6.5 Pg of SOC is stored in the top 1 m of the soils in Nigeria. In addition, about 48%, 27% and 25% of the total SOC stock in the Nigerian soils is found at 0-30 cm, 30-60cm and 60-100cm depth respectively (See Table 4.5). For the sake of brevity, further presentation of results on SOCD and SOCS focuses only on the best prediction model (RFM).

4.3.4 SOC density and total stock under various land use types

Mean SOCD and total SOCS differs in soils under different land use. Soils under FL have a higher SOCD (99.2 Mg C ha^{-1}) while soils under grassland have the lowest SOCD (51.0 Mg C ha^{-1}(see Table 4.5). In the top 30 cm, soils under savannah and cropland presents similar mean SOCD. However, soils under savannas LULC have a higher SOCD below 30 cm soil depth. SOCS values ranged from 4.2 Tg to 662 Tg with highest storage in the 30-60 cm soil layer (Table 4.5). SOCS distribution across the various land use types followed a slightly different trend as SOCD (Fig. 4.8) with SOCS distributing deeper in soils of the savannas (38 %) and shallower in soils of shrublands (1.1%). The distribution of the SOCS considering the percent total storage follow the trend S (38.4 %) > CL (30.5%) > FL (16.1%) > GL (13.3%) >SL (1.1%). This trend varied from one AEZ to another (see Fig. 4.9). Except for the SDS and SHS AEZs, savannas, forestlands and croplands ranked top in terms of total SOC stored across the study area.

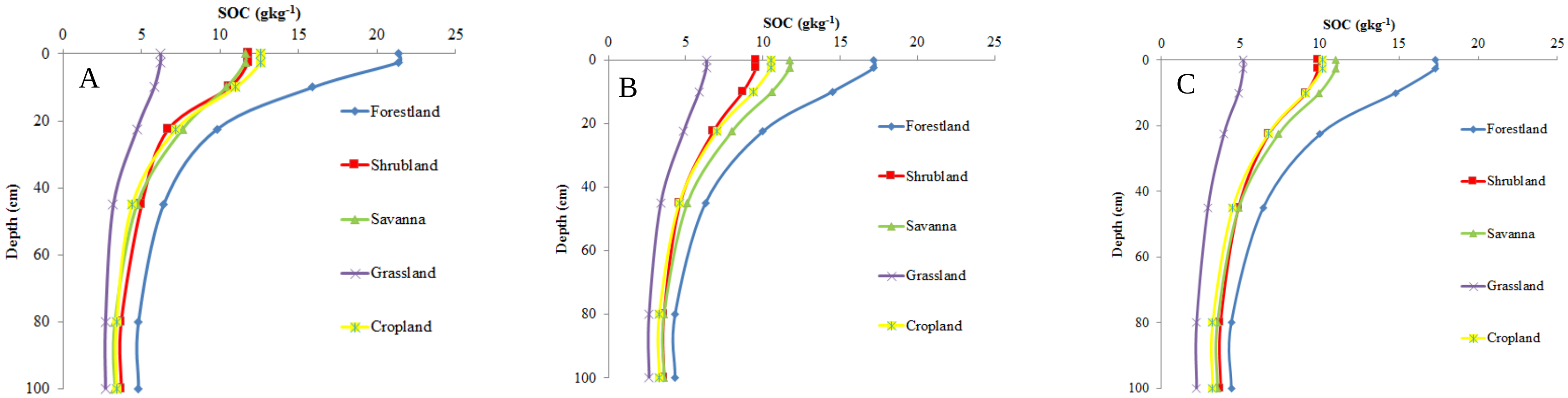


Figure 4.5. Vertical distribution of SOC concentration (grams per kilogram) in the soil profile under different land use types as predicted by spline functions (A), Random Forest model (B) and Cubist Model (C).

Table 4.5. Mean and standard deviation of SOC density and total SOC stock in each land use type and agro-ecological zone.

		SOC Stock (Mg ha^{-1})					Total SOC storage (Tg)				
		Depth (cm)									
Landuse	Area (10^6 m^2)	0-5	5-15	15-30	30-60	60-100	0-5	5-15	15-30	30-60	60-100
Forest	102996	10.9±4.5	18.9±7.5	20.1±7.3	25.7±9.0	23.5±7.5	113.2	194.6	207.0	264.4	242.3
Shrubland	8576	6.3±2.5	11.5±4.6	13.6±4.9	18.4±5.5	18.8±4.8	5.4	9.9	11.6	15.8	16.1
Savanna	326279	7.3±2.4	13.4±4.2	15.3±4.4	19.7±4.9	18.2±4.6	239.7	435.9	500.0	642.3	594.3
Grassland	166359	4.4±1.3	8.7±2.4	10.1±2.5	14.1±2.9	14.2±2.6	73.2	135.9	168.5	234.0	236.4
Cropland	290970	6.9±2.7	12.4±4.7	14.1±4.3	18.5±4.6	17.5±4.0	201	361.7	409.2	538.1	508.7
Agro-ecological zone											
Humid Forest	106184	13.3±2.9	23.1±5.3	23.7±5.8	30.2±7.0	27.1±6.2	141.7	245.7	252.1	320.8	287.9
Derived Savannah	258200	8.1±2.7	14.6±4.5	16.5±4.7	20.8±4.9	18.9±4.7	208.2	375.6	427.3	536.8	486.9
Southern Guinea Savannah	144622	6.2±1.3	11.3±2.3	13.2±2.7	17.6±3.2	16.9±3.2	89.0	163.2	191.4	254.0	244.0
Northern Guinea Savannah	113377	6.2±1.0	11.1±1.7	12.8±1.8	17.0±2.1	16.5±2.3	69.8	125.6	145.3	191.4	187.0
Sudan Savannah	174997	5.0±0.9	9.3±1.5	11.3±1.6	15.4±1.9	115.0±1.9	88.1	162.0	197.9	268.8	261.8
Sahel Savannah	90987	3.9±1.2	7.3±2.0	9.1±2.1	13.14±2.5	14.0±3.1	35.7	66.0	82.8	118.8	127.2
Mid High Altitude	20436	7.4±1.5	13.0±2.6	14.8±3.1	19.8±3.4	19.2±3.0	15.1	26.6	30.2	40.5	39.3

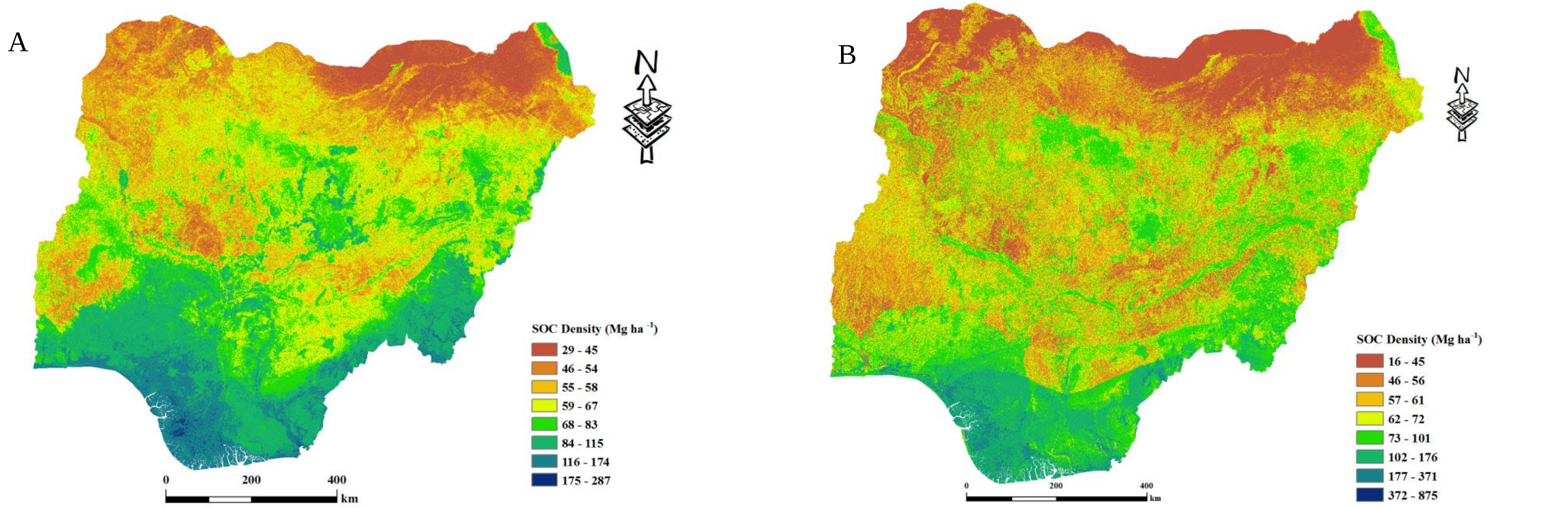


Figure 4.6. Spatial distribution of the mean SOC density (Mg ha-1) in the top 1m of soil as predicted by Random forest model (A) and Cubist (B).

4.3.5 SOC density and total stock across AEZs

SOCD varied significantly across the AEZs (Table 4.5). It ranged between 47.3 and 117.6 Mg C ha^{-1}. There was more SOCD in soils under HF than in other AEZs. Overall, SOCD in the 1m depth across the various AEZs includes HF (117.6 Mg C ha^{-1}), DS (78.8 Mg C ha^{-1}), MA (74.2 Mg C ha^{-1}), SGS (65.1 Mg C ha^{-1}), NGS (63.6 Mg C ha^{-1}), SDS (55.9 Mg C ha^{-1}) and SHS (47.3 Mg C ha^{-1}). Also, the distribution of total SOC stored in soils under the various AEZs followed the sequence DS (31.7%) > HF (18.3%) > DS (15.2%) > SGS (14.6%) > NGS (11.1%) > SHS (6.7%) > MA (2.4%) (see Fig 4.8).

4.3.6 SOC density and total stock under various soil types

Mean SOCD differs across the various soil types (see Table 4.6) with values ranging from 4.8 to 25.9 Mg C ha^{-1}. In the top 30 cm, Ferralsols showed slightly higher SOCD than other soil types. However, Gleysols have the highest SOCD below 30 cm. Ferralsols and Gleysols collectively have higher Mean SOCD (109 Mg C ha^{-1}) at the top 1m soil depth while Arenosols have the lowest mean SOCD (53.8 Mg C ha^{-1}). In terms of total SOC storage to 1 m depth, Lixisols (1510.6 Tg) show the highest while Phaeozems (22.40 Tg) show the lowest total SOC stock (Table 4.7).

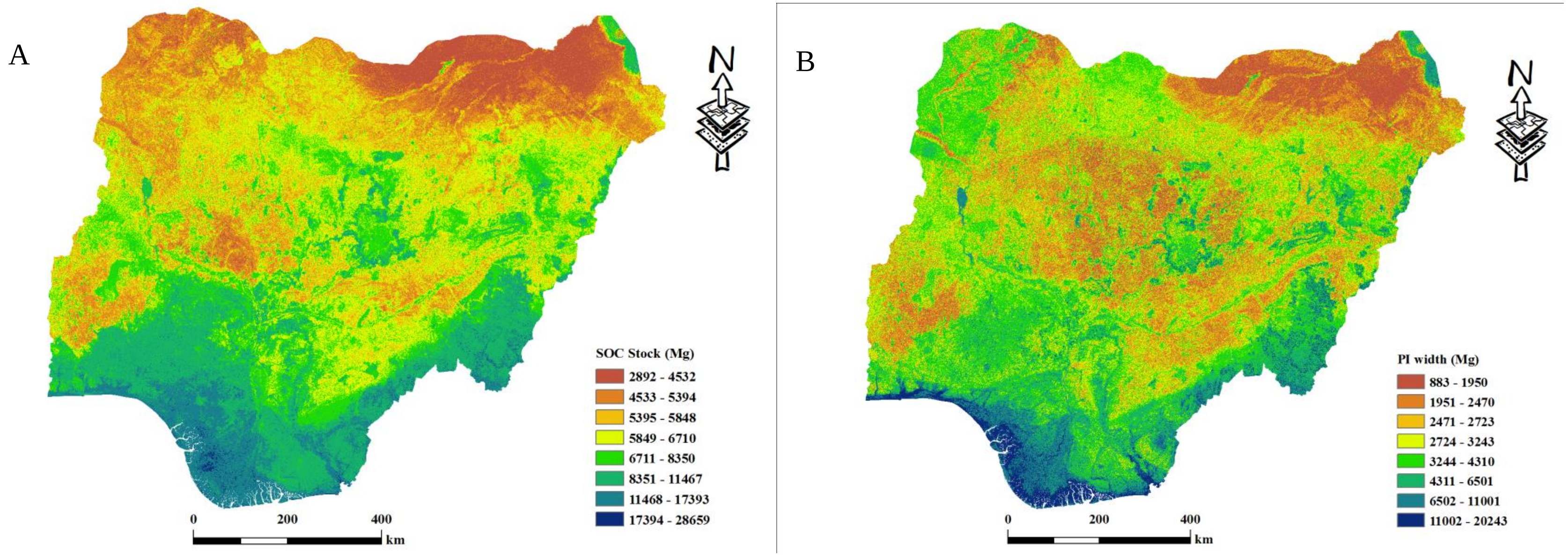


Figure 4.7 Spatial distribution of predicted mean SOC stock (Mg) based on Random forest model (A) and the associated 95% prediction interval (B) in the top 1m of soil.

Table 4.6. Mean and standard deviation of SOC density and total SOC stock of major WRB soil groups.

	Area (10^6 m^2)	SOC Stock (Mg ha^{-1})					Total SOC storage (Tg)				
		Depth (cm)									
Soils		0-5	5-15	15-30	30-60	60-100	0-5	5-15	15-30	30-60	60-100
Acrisols	40241	7.8±2.7	13.9±4.5	15.8±4.8	20.2±5.2	18.5±4.5	76.1	31.2	55.9	63.5	74.4
Arenosols	151599	4.8±1.9	8.8±3.5	10.9±3.6	14.8±3.9	14.6±3.4	53.8	72.1	133.7	164.5	221.0
Cambisols	3662	7.1±1.8	12.6±3.1	14.4±2.7	18.6±3.1	18.0±2.5	70.8	2.6	4.6	5.3	6.6
Ferralsols	13132	12.6±2.4	21.9±4.3	21.8±4.7	28.6±5.2	24.9±4.9	109.8	16.6	28.7	28.6	32.7
Fluvisols	62323	6.7±4.0	12.2±7.1	14.4±7.4	19.2±8.7	18.3±7.2	70.8	41.9	76.3	89.6	114.3
Gleysols	30865	11.9±5.3	21.0±9.1	21.7±8.8	28.6±11.5	25.9±9.0	109.0	36.7	64.8	66.9	79.8
Leptosols	113962	7.2±2.5	13.0±4.3	15.0±4.7	19.4±4.9	18.4±4.4	72.9	82.1	147.8	170.6	209.4
Lixisols	217529	6.9±2.1	12.4±3.4	14.2±3.5	18.5±3.9	17.4±3.7	69.4	149.6	269.2	309.3	379.1
Luvisols	48442	6.2±1.0	11.5±1.8	13.6±2.1	17.8±2.9	16.5±3.4	65.6	30.2	55.7	65.7	79.8
Nitisols	129900	9.5±3.8	16.7±6.3	17.8±6.0	22.4±6.5	20.8±5.6	87.2	123.3	216.2	231.5	270.5
Phaeozems	2920	7.3±1.1	13.1±2.0	15.2±2.1	20.4±2.9	20.7±3.0	76.7	2.1	3.8	4.4	6.0
Plinthosols	49074	6.2±1.3	11.3±2.3	13.3±2.6	17.2±2.7	15.4±2.5	63.4	30.2	55.5	65.2	75.6
Solonchaks	5038	5.6±0.9	10.1±1.6	12.0±1.7	16.3±1.7	17.3±2.2	61.3	2.8	5.1	6.1	8.7
Vertisols	14678	5.6±1.7	10.5±3.2	12.6±2.9	17.6±2.4	17.8±2.5	64.0	8.3	15.3	18.4	26.1

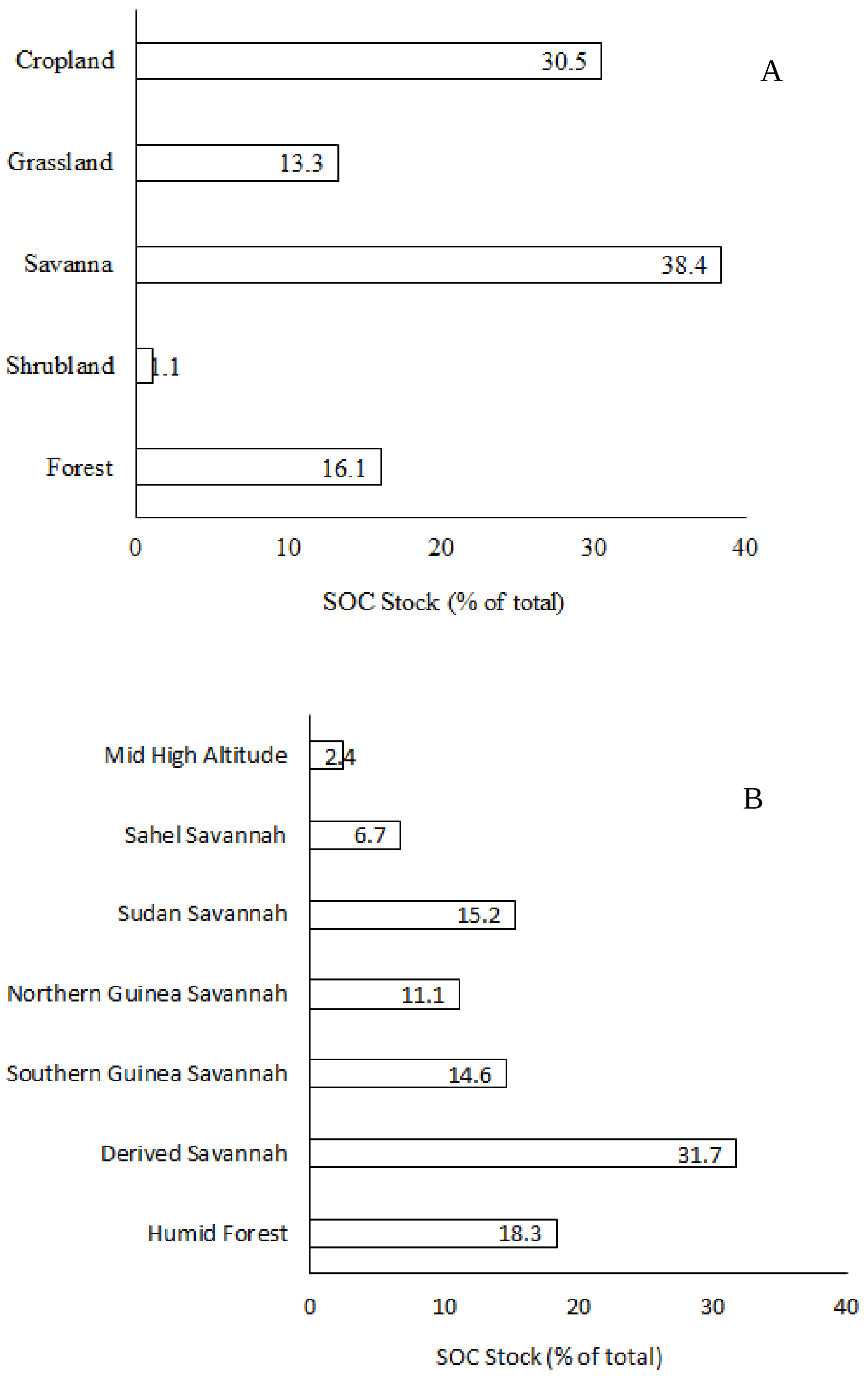


Figure 4.8. Percentage of SOC stored at 1 meter soil layer under different land use types (A) and agro-ecological zones (B) in Nigeria.

4.3.7 Uncertainty in Total SOC stock estimation

The results of the 95% prediction interval of SOC stock estimated based on RFM is presented in Fig. 4.7 while Table 4.7 shows the distribution of total SOC stock and their associated uncertainty across the various LULC, AEZ and soil groups. A closer look at Fig. 4.7 will reveal a wide prediction interval around the Delta areas in the southern fringes of the country, the borders of the north-eastern region as well as the north-central region. Among the various land use, there is a wider uncertainty in the estimated SOCS under the Savanna (±6.00) than other land use types. Also, the DS agro-ecological zone shows the widest range of uncertainty (±4.91 Tg) while there is higher uncertainty in our SOCS estimation for Lixisols (±3.46 Tg) compared to other soil groups (Table 4.7).

4.3.8 Potential soil carbon sequestration

The potential soil C sequestration under different LULC types across the various AEZs of Nigeria is presented in Fig. 4.10. On average the potential to sequester SOC ranges from -17.0 and 30.8 Mg C ha^{-1} depending on the LULC change and AEZ (See Fig. 4.10a). The DS, which is transitional between rainforest and savannas, has the highest capacity to store C (19.0 to 30.8 Mg C ha^{-1}) especially with the restoration of shrublands to forests (30.8 Mg C ha^{-1}) while the Southern Guinea Savannah (0.4 to 2.3 Mg C ha^{-1}) has the least capacity to store additional C. With the exception of the HF and DS zones, soils under grasslands show the highest potential to sequester C. This is followed by shrublands and croplands. In the HF zone, C sequestration ranged from 3.8 to 22.8 Mg C ha^{-1} with an average of 16.9 Mg C ha^{-1}. The restoration of shrublands, croplands, grasslands, and savannas to the pristine vegetation in this AEZ has the potential of storing additional 3.8, 19.9 Mg C ha^{-1}, 21.1 Mg C ha^{-1} and 22.8 Mg C ha^{-1} respectively. These values represent a change of about 3.1%, 18.6%, 19.9% and 21.9% between the SOC in the current land use and the native vegetation (see Fig. 4.10b).

In the DS AEZ, the restoration of the various land use types gave soil C sequestration in the range of 19.0 to 30.8 Mg Cha^{-1} with an average of 25.5 Mg C ha^{-1}. The restoration of various LULC types to forest has the potential of storing additional 30.8 Mg C ha^{-1} (shrublands), 29.2 Mg C ha^{-1} (grasslands), 23.1 Mg C ha^{-1} (savannas) and 19.0 Mg C ha^{-1} (croplands). These reflect a change of about 44.8%, 41.5%, 30.2% and 23.5 % between SOC under these land use types and forest vegetation in the DS agro-ecosystem.

Across the various Savannah AEZs, C sequestration potential under the different land use types is as follows SHS (-17.0 to 19.0 Mg C ha^{-1}) > SDS (0.4 to 6.9 Mg C ha^{-1}) > NGS. (0.2 to 3.4 Mg C ha^{-1}) > SGS (0.4 to 2.3 Mg C ha^{-1}). The restoration of grasslands has the highest potential to sequester additional C (1.6 to 19.0 Mg C $ha^{-}1$) compared to other land use types these AEZs. However, the restoration of forestlands in the SHS leads to a reduction in C sequestration (-17.0 Mg C ha^{-1}) and is indicative of a loss of C from the system. This could be explained by the high biomass C input from exotic forest plantations in some parts of the SHS agro-ecosystem. These are in the form of shelterbelt or agroforestry projects.

Table 4.7. Mean SOC Stocks, their uncertainty and Total SOC stocks in the top 1 m.

	Mean (Mg ha^{-1})	SD (Mg ha^{-1})	Total SOC Stock (Tg)	95% Confident limit (Tg)
Soils				
Acrisols	76.1	9.9	306.2	±0.78
Arenosols	53.8	7.4	815.6	±2.21
Cambisols	70.8	6	25.9	±0.06
Ferralsols	109.8	9.9	144.1	±0.42
Fluvisols	70.8	15.8	441.5	±1.15
Gleysols	109	20.1	336.4	±1.06
Leptosols	72.9	9.5	830.9	±2.15
Lixisols	69.4	7.5	1510.6	±3.46
Luvisols	65.6	5.4	317.5	±0.75
Nitisols	87.2	12.8	1132.6	±2.84
Phaeozems	76.7	5.3	22.4	±0.06
Plinthosols	63.4	5.3	311	±0.73
Solonchaks	61.3	3.8	30.9	±0.08
Vertisols	64	5.8	94	±0.21
Landuse				
Forest	99.2	16.3	1021.5	±2.96
Shrubland	68.6	10.2	58.8	±0.17
Savanna	73.9	9.4	2412.2	±6.00
Grassland	51	5.3	847.9	±2.26
Cropland	69.4	9.2	2018.6	±4.95
Agro-ecological zone				
Humid Forest	117.6	12.6	1248.2	±3.8
Derived Savannah	78.8	9.8	2034.7	±4.91
Southern Guinea Savannah	65.1	5.9	941.7	±2.33
Northern Guinea Savannah	63.6	4.1	720.6	±1.63
Sudan Savannah	55.9	3.6	978.6	±2.72
Sahel Savannah	47.3	5.1	430.4	±1.09
Mid High Altitude	74.2	6.3	151.7	±0.35

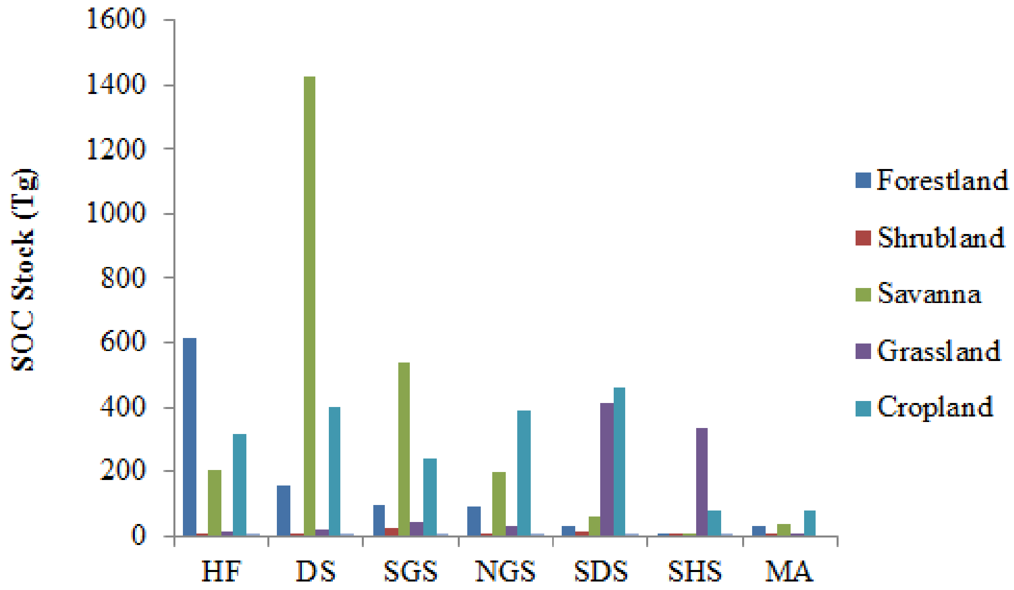


Figure 4.9. Total SOC (Tg) stored in 1 m soil depth under different land use types across various agro-ecological zones of Nigeria. Abbreviations: HF; Humid forest, DS; Derived savannah, SGS; Southern Guinea savannah, NGS; Northern Guinea savannah, SDS; Sudan savannah, SHS; Sahel savannah, MA; Mid-High Altitude.

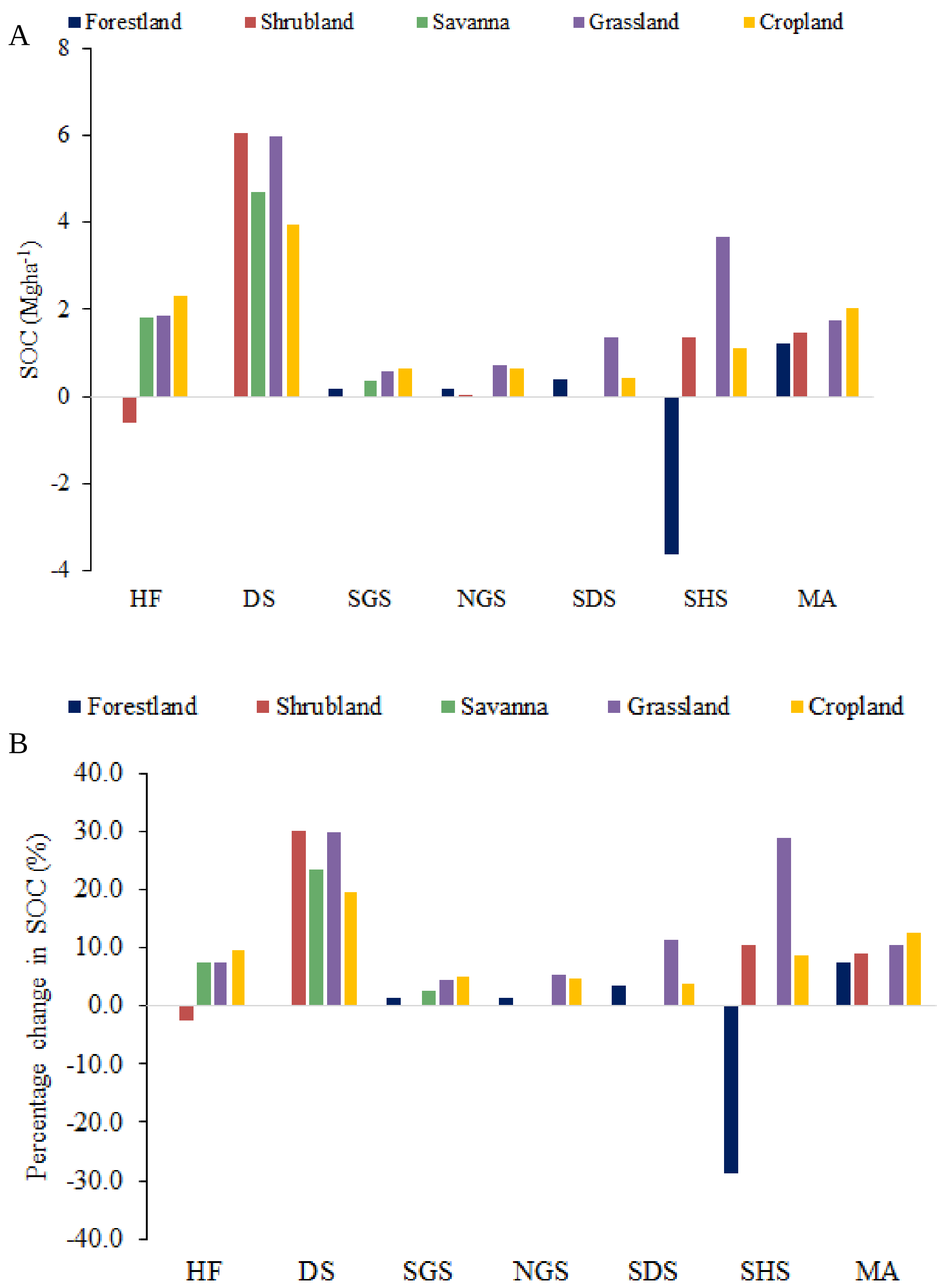


Figure 4.10. Potential Carbon Sequestrability (A) and percentage change in SOC (A) of selected land use types across various agro-ecological zones. Abbreviations: HF; Humid forest, DS; Derived savanna, SGS; Southern Guinea savanna, NGS; Northern Guinea savanna, SDS; Sudan savanna, SHS; Sahel savanna, MA; Mid-High Altitude.

4.4 Discussion

4.4.1 Modelling SOC concentration

Our results indicate that RFM and Cubist exhibited similar but better performance than BRT in predicting SOC and bulk density (Table 4.3). This is in contrasts to the reports of Yang et al (2016) who compared the performances of RFM and BRT in predicting SOC. Our reported Pc based on the three models used for SOC prediction is relatively low (0.15 to 0.48), indicating a low to moderate agreement between the predicted and measured SOC, with the strongest agreement obtained for the upper 0–15 cm soil depth. The agreement weakens as we go down the profile with increasing systematic deviations from the 45° line exhibited by layers below 30cm depth. This is expected since soil is less variable at the lower depth, which corresponds with a more natural state of the soil than the surface layer. SOC is sensitive to soil perturbations and as such has a higher variability in the soil surface than the sub-surface. Our reported Pc values for RFM (Table 4.3), which was the best performing model, are comparable to those reported by Malone et al. (2009) but better than the values reported by Lacoste et al. (2014) and Mulder et al. (2016). In contrast, Yang et al (2016) reported a better performance by RFM for SOC predicted in an alpine ecosystem using similar approaches. These dissimilar performances could be attributed to the difference in the sources of data, scale of prediction and types of predictors (Miller et al., 2015). The better performance of RFM in terms of RMSE could be ascribed to its better capabilities in dealing with non-linear and hierarchical relationships between SOC and environmental variables.

Our study revealed that soil type, climate, vegetation indices and terrain attributes are important predictors of SOC. This corroborates reports from previous studies (Jobbagy and Jackson, 2000; Albaladejo et al., 2013). Several studies have highlighted the importance of climate in predicting SOC contents at the regional, national and continental scales (Wynn et al., 2006; Rusco et al., 2001; Martin et al., 2011; Meersmans et al., 2011; 2012; Adhikari et al., 2014). At the supra-national to continental scale, Rusco et al. (2001) reported that SOC is positively correlated with precipitation amount and negatively correlated with average temperature. This was also found by Martin et al. (2011) and Meersmans et al. (2011) who highlighted a stronger correlation between SOC and precipitation than with temperature. Precipitation plays a key role in biomass productivity which determines litter input to the soil (Chaplot et al., 2010) while temperature influences the decomposition of litter in the soil as well as C mineralization rate.

4.4.2 SOC density and stock

The average SOC density reported in this study is within the predicted range of SOC density reported for Senegal but slightly above the value for other soils in West African (Batjes, 2001). It is however below the range of 7.6-7.7 kg C m^{-2} reported for soils in the warm savannah region of Central Africa (Batjes, 2008). The total SOC stock of 6.5 Pg reported in this study represents 4% of the estimate of total SOC stock at 0-100cm soil depth for the entire Africa and is slightly above the 5.1 Pg estimated by Henry et al. (2009) for the same study area. This study revealed that about 48%, 27% and 25% of the SOC storage is in the 0-30 cm, 30-60 cm and 60-100 cm depth intervals respectively. This is lower than the 70% C reported by Albaladejo et al. (2013) in the top 40 cm of soils in semi-arid areas of Spain but consistent with the 45% C reported by Batjes (2008) in the top 30 cm of soils in Central Africa. Our results also corroborate recent work by Mulder et al. (2016) using a similar approach.

The SOC density and stock results showed a gradual increase with depth. This is consistent with previous studies (Adhikari et al., 2014; Dorji et al., 2014; Bonfatti et al., 2016; Mulder et al., 2016) and is partly due to the effect of the depth interval (e.g. 40 cm for the 60-100cm interval compared to 5cm for the 0-5cm depth interval). Although there were higher SOC concentrations in the surface layers than sub-surface, the differences in the width of depth intervals is enough to offset it. Also, the difference between the trend in SOC density and stock across the AEZs could be attributed to the variations in land area covered by the various AEZs. The observed trends in the spatial distribution of SOCD and SOCS could be attributed to a variety of factors especially precipitation, soil texture and temperature. In Nigeria, annual rainfall decreases from about 2500 mm in the south-western areas bordering the Atlantic Ocean to about 500 mm in the north-eastern areas bordering Lake Chad (Akpa et al., 2014). Previous studies have shown a significant relationship between SOC stock and precipitation (Post et al., 1982; Dixon et al., 1994), soil texture (Borchers and Perry, 1992) as well as temperature (Nishina et al., 2014). Borchers and Perry (1992) reported a lower SOC concentration with coarser soils than fine textured soils. Moreso, Burke et al. (1995) reported that organic C increased with precipitation and clay content, and decreased with temperature.

4.4.3 Influence of land use on SOC

Our SOC concentrations decreased with depth and vary significantly across all land use types. This difference among land use types could be attributed to differences in the proportion of SOC contributed by the biomass of the various vegetation types (Young et al., 2005; Obalum et al., 2012) and the varying level of soil perturbation associated with the different LULC types (Six et al., 2002). The higher SOC concentration in the forests and shrublands compared to grasslands could be attributed to the higher aboveground biomass under these land uses compared to grasslands (Martin et al., 2010; Dorji et al., 2014; Wasige et al., 2014). Higher aboveground organic material input and relatively low rates of decomposition have been associated with increased SOC levels in forests compared to grasslands (Jobbágy and Jackson, 2000; Guo and Gifford, 2002; Don et al., 2011; Bonfatti et al., 2015).The higher SOC concentration of croplands compared to grasslands in our study area could be attributed to effect of overgrazing and burning of grasslands by pastoralist in the northern part of the country which has large grasslands coverage. Burning is a common practice by pastoral farmers in Nigeria especially during the dry harmattan season to allow regeneration of grasses for animal feeds. The resultant increase in soil temperature can reduce SOC due to increased soil organic matter mineralization (Ando et al., 2014) while the loss of vegetation cover due to incessant burning will further result in low SOC through soil erosion. Although the net SOC loss through soil erosion across the entire country is negligible due to the balance effect of soil redistribution (Ritchie et al., 2007), at the various agro-ecological zone where there is relatively uniform soil type, landform, cropping system, and climatic factors favouring erosion, soil erosion is an important process controlling the levels of C in the soil (Gregorich et al., 1998).

4.4.4 Uncertainty in SOC stock estimation

There is a wide range of uncertainty in our estimation of SOCS which could be attributed to the below optimal profile datasets used in this study as well as inherent uncertainties in the sources of data used. In this study we have used legacy soil data. Although legacy data generally presents opportunities for DSM in data-scarce countries of the world (Mayr et al., 2008), the use of such legacy data poses serious challenges due to uneven spread and age of the data (Krol, 2008). The uneven spread of data across Nigeria (Fig. 4.2) could be responsible for the wide prediction interval of the estimated SOC stock reported here (Fig. 4.7). Similarly, in the case of each of the land use types and AEZ, the uneven distribution of data could have impacted on the accuracy of the predicted variables as the some of the prediction values could be outside the range of values of the training subset. Future sampling scheme will need to target those areas with wide prediction uncertainty to improve upon the accuracy of our prediction.

4.4.5 Potential soil carbon sequestration

There is potential for SOC sequestration in the study area considering the differences in SOC stock across the prevailing land use types and AEZ (Fig. 4.9). The DS has the highest capacity to store more C especially with the restoration of shrublands and grasslands to forest plantation (Fig. 4.10). This could be explained by the high rate of human interference in the native forests of this zone manifested by indiscriminate tree logging for farming, housing and energy. In Nigeria, annual loss of forest has been put at approximately 3,500 km2 (Ravilious et al., 2010). Most of these losses occur in the densely populated southern part of the country where this AEZ is located. SOC content is dependent upon soil organic matter inputs from plant biomass and root turnover, thus the destruction of vegetation will reduce biomass inputs, leading to the exposure of soil surface and increased SOC decomposition.

The relatively high C sequestration capacity of soils under the SHS is a reflection of the degraded nature of soils under this AEZ (Raji and Ogunwole, 2006). Both environmental and anthropological factors including; low precipitation rate, increased soil temperature due to incessant vegetation burning, heavy grazing and coarse nature of soils have effect on C accumulation in soils. Low rainfall and reduced length of growing period could result in poor biomass yield and low accumulation of soil densities especially in semi-arid areas (Sreenivas et al., 2014). Heavy grazing has been reported to have a damaging effect on shrubs and trees, and grasslands (McIvor et al., 1995). Howden et al. (1999) reported a large difference in carbon sequestration between an ungrazed-never burnt and a grazing-annual burning grasslands in Australia. This was attributed to the negative effect of burning on woody biomass. The restoration of forestlands in the SHS could lead to a reduction in C sequestration (-17.0 Mg C ha^{-1}) and is indicative of C loss from the system. This could be explained by the high biomass C input from the increasing extent of exotic tree plantations in the form shelter belts from desert encroachment mitigation projects (Adegbehin and Igboanugo, 1990; Adegbehin et al., 1990; Verinumbe, 1991). Another reason could be that the number of SOC data in this AEZ was not enough to capture the C variation between the pristine vegetation and forest land use.

4.5 Conclusions

We estimated the total soil organic carbon and stocks for soils under different land use types across different agro-ecological zones of Nigeria. Based on the overall analysis, the following conclusions can be drawn:

- Soil type, climate, vegetation indices and terrain attributes are important predictors of SOC for Nigeria.
- Mean SOC concentration ranged between 7.4 and 11.0 g C kg^{-1} in the topsoil (0-30cm depth) and between 2.8 and 4.8 g C kg^{-1} in the subsoil (30-100cm depth)
- Forest LULC, Humid Forest AEZ and Ferralsols have the highest SOC density while grasslands, Sahel Savannah and Arenosols have the lowest SOC density.
- Total SOC stored in the top 1 m in Nigeria was 6.5 Pg with an average density of 71.60 Mg C ha^{-1}. This represents 4% of the estimated total SOC stock in the top 0-100cm soil reported previously for the entire Africa.
- SOC density and stock varies greatly across the study area; decreasing from the southwest to the northeast, and increasing from Sahel Savannah to Humid Forest agro-ecological zones.
- Restoration of the various land use types to their natural ecosystem has the potential to sequester about 0.2 to 30.8 Mg C ha^{-1} depending on the LULC and AEZ.
- The Derived Guinea Savannah presents a potential hotspot for targeted carbon sequestration projects in Nigeria.

References

Adegbehin, J. O. & Igboanugo, A. B. I. 1990. Agroforestry practices in Nigeria. *Agroforestry Systems*, *10*(1), 1-22.

Adegbehin, J. O., Igboanugo, A. B. I., & Omijeh, J. E. 1990. Potentials of agroforestry for sustainable food and wood production in the savanna areas of Nigeria. *Savanna*, *11*(2), 12-26.

Adhikari, K., Hartemink, A. E., Minasny, B., Kheir, R. B., Greve, M. B., & Greve, M. H. 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. *PloS ONE*, *9*(8), e105519.

Akpa, S. I.C., Odeh, I. O.A, Bishop, T. F., & Hartemink, A. E. 2014. Digital mapping of soil particle-size fractions for Nigeria. *Soil Science Society of America Journal*, 78, 1953-1966. DOI:10.2136/sssaj2014.05.0202.

Albaladejo, J., Ortiz, R., Garcia-Franco, N., Navarro, A. R., Almagro, M., Pintado, J. G., & Martínez-Mena, M. 2013. Land use and climate change impacts on soil organic carbon stocks in semi-arid Spain. *Journal of Soils and Sediments*, *13*(2), 265-277.

Ando, K., Shinjo, H., Noro, Y., Takenaka, S., Miura, R., Sokotela, S. B., & Funakawa, S. 2014. Short-term effects of fire intensity on soil organic matter and nutrient release after slash-and-burn in Eastern Province, Zambia. Soil Science and Plant Nutrition, 60(2), 173-182.

Anikwe, M.A.N. 2010. Carbon storage in soils of southeastern Nigeria under different management practices. *Carbon Balance and Management* **5** (1), 1-7

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J. Mendonça Santos, M.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., & Zhang, G. L. 2014. Chapter Three-GlobalSoilMap: Toward a fine-resolution global grid of soil properties. *Advances in Agronomy*, *125*, 93-134.

Batjes, N.H. 2001. Options for increasing carbon sequestration in West African soils: an exploratory study with special focus on Senegal. *Land Degradation and Development*, 12, 131–142.

Batjes, N. H. 2008. Mapping soil carbon stocks of Central Africa using SOTER. *Geoderma*, *146*(1), 58-65.

Bekele, L., Kleja, D.B., Nilsson, I., & Olsson, M. 2006. Soil carbon sequestration under different exotic tree species in the southwestern highlands of Ethiopia. *Geoderma*, 136, 886–898.

Bishop, T. F. A., McBratney, A. B., & Laslett, G. M. 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, *91*(1), 27-45.

Bonfatti, B. R., Hartemink, A. E., Giasson, E., Tornquist, C. G., & Adhikari, K. 2016. Digital mapping of soil carbon in a viticultural region of Southern Brazil. *Geoderma*, 261, 204-221.

Borchers, J. G., & Perry, D. A. 1992. The influence of soil texture and aggregation on carbon and nitrogen dynamics in southwest Oregon forests and clearcuts. *Canadian Journal of Forest Research*, *22*(3), 298-305.

Breiman, L. 2001a. Random forests. *Machine Learning*, 45(1), 5–32.

Bruce, J. P., Frome, M., Haites, E., Janzen, H., Lal, R., & Paustian, K. 1999. Carbon sequestration in soils. *Journal of Soil and Water Conservation*, *54*(1), 382-389.

Bui, E., Henderson, B., & Viergever, K. 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. *Global Biogeochemical Cycles*, *23*(4).

Burke, I. C., Elliott, E. T., & Cole, C. V. 1995. Influence of macroclimate, landscape position, and management on soil organic matter in agroecosystems. *Ecological Applications*, 124-131.

Chaplot, V., Bouahom, B., & Valentin, C. 2010. Soil organic carbon stocks in Laos: spatial variations and controlling factors. *Global Change Biology*, *16*(4), 1380-1393.

Collard, F., Saby, N. P. A., Richer-de-Forges, A. C., Lehmann, S., Paroissien, J. B., & Arrouays, D. 2014. Spatial prediction of soil organic carbon at different depths using digital soil mapping. In: Arrouays, D., McKenzie, N., Hempel, J., Richer-de-Forges, A.C., McBratney, A. B. (Eds.). GlobalSoilMap: Basis of the global spatial soil information system, CRC press. pp. 181-184.

Del Galdo, I., Six,J., Peressotti, A., & Cotrufo, M.F. 2003. Assessing the impact of land-use change on soil C sequestration in agricultural soils by means of organic matter fractionation and stable isotopes. *Global Change Biology*, 9, 1204–1213

Dikau, R., Brabb, E.E. & Mark, R.M. 1991. Landform classification of New Mexico by computer. Open File Rep. 91-634. USGS, Denver, CO.

Dixon, J. A., Gibbon, D. P., & Gulliver, A. 2001. Farming systems and poverty: improving farmers' livelihoods in a changing world. Food and Agricultural Organization (FAO). Rome, Italy. 412p

Dixon, R.K., Solomon, A.M., Brown, S., Houghton, R.A., Trexier, M.C., & Wisniewski, J. 1994. Carbon pools and flux of global forest ecosystems. *Science*, 263, 185–190.

Don, A., Schumacher, J., & Freibauer, A. 2011. Impact of tropical land-use change on soil organic carbon stocks–a meta-analysis. *Global Change Biology*, *17*(4), 1658-1670.

Dorji, T., Odeh, I. O., Field, D. J., & Baillie, I. C. 2014. Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems, Eastern Himalayas. *Forest Ecology and Management*, *318*, 91-102.

Elith, J., Leathwick, J. R., & Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802-813.

Fang, X., Xue, Z.J., Li, B.C., & An, S.S. 2012. Soil organic carbon distribution in relation land use and its storage in a small watershed of the Loess Plateau, China. *Catena*, 88, 6–13. doi:10.1016/j.catena.2011.07.012

FMANR, 1990. Literature Review on Soil Fertility Investigations in Nigeria. Federal Ministry of Agricultural and Natural Resources, Lagos, Nigeria, 281p.

Follett, R. F. 2001. Soil management concepts and carbon sequestration in cropland soils. *Soil and Tillage Research*, *61*(1), 77-92.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., & Huang, X. 2010. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, *114*(1), 168-182.

Giller, K. E., Corbeels, M., Nyamangara, J., Triomphe, B., Affholder, F., Scopel, E., & Tittonell, P. 2011. A research agenda to explore the role of conservation agriculture in African smallholder farming systems. *Field crops research*, *124*(3), 468-472.

Goovaerts, P. 2001. Geostatistical modelling of uncertainty in soil science. *Geoderma* 103 (1– 2): 3–26.

Gregorich, E. G., Greer, K. J., Anderson, D. W., & Liang, B. C. 1998. Carbon distribution and losses: erosion and deposition effects. *Soil and Tillage Research*, *47*(3), 291-302.

Gregorich, E. G., Rochette, P., VandenBygaart, A. J., & Angers, D. A. 2005. Greenhouse gas contributions of agricultural soils and potential mitigation practices in Eastern Canada. *Soil and Tillage Research*, *83*(1), 53-72.

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis. *Geoderma*, *146*(1), 102-113.

Gulledge, J., & Schimel, J. P. 2000. Controls on soil carbon dioxide and methane fluxes in a variety of taiga forest stands in interior Alaska. *Ecosystems*, *3*(3), 269-282.

Guo, L. B., & Gifford, R. M. 2002. Soil carbon stocks and land use change: a meta-analysis. *Global Change Biology*, *8*(4), 345-360

Henry, M., Valentini, R., & Bernoux, M. 2009. Soil carbon stocks in ecoregions of Africa. *Biogeosciences Discussions*, *6*(1), 797-823.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*. 25, 1965–1978.

Howden, S. M., McKeon, G. M., Walker, L., Carter, J. O., Conroy, J. P., Day, K. A., Hall, W. B., Ash, A. J & Ghannoum, O. 1999. Global change impacts on native pastures in south-east Queensland, Australia. *Environmental Modelling & Software*, *14*(4), 307-316.

IITA (International Institute *of* Tropical Agriculture), 1992. *Medium Term Plan 1994-1998,* IITA, Ibadan, Nigeria. 136 p.

IPCC, 2011. IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation In: O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, S. Schlömer, C. von Stechow (eds), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 1076p.

Iwahashi, J., & Pike, R. J. 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*, *86*(3), 409-440.

Jagtap, S.S. 1994. Resource Analysis and Information System, RIS, User's Guide, Agro-ecological Studies Unit, IITA, Ibadan, Nigeria.

Jagtap, S.S. 1995. Environmental characterization of the moist lowland savanna of Africa. In: Kang, B.T. (Ed.). Moist savannas of Africa: pontentials and constraints for crop production. Proceedings of an IITA/FAO workshop held from 19-23 September, 1994. Cotonou, Republic of Benin. p 9-30.

Jalabert, S. S. M., Martin, M. P., Renaud, J. P., Boulonne, L., Jolivet, C., Montanarella, L., & Arrouays, D. 2010. Estimating forest soil bulk density using boosted regression modelling. *Soil Use and Management*, *26*(4), 516-528.

Jenny, H. 1980. The Soil Resource: origin and behavior Springer-Verlag, New York.

Jobbágy, E. G., & Jackson, R. B. 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological applications*, *10*(2), 423-436.

Keay, R. W. J. 1959. Vegetation map of Africa south of the Tropic of Cancer. Explanatory notes. *Vegetation map of Africa south of the Tropic of Cancer. Explanatory notes*.

Krol, B. G. C. M. 2008. Towards a data quality management framework for digital soil mapping with limited data. In *Digital Soil Mapping with Limited Data* (pp. 137-149). Springer Netherlands.

Kuhn, M., Weston, S., Keefer, C., Quinlan, N.C.C. 2013. C code for Cubist by Ross Quinlan 2013. Cubist: Rule- and Instance-Based Regression Modeling., R package version

0.0.13.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., & Walter, C. 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. *Geoderma*, *213*, 296-311.

Lal, R. 2005. Forest soils and carbon sequestration. *Forest Ecology and Management*, *220* (1), 242-258.

Lawrence, R., Bunn, A., Powell, S., Zambon, M. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90 (3), 331–336.

Leenaars, J.G.B. 2012. Africa Soil Profiles Database, version 1.0.: A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). ISRIC Rep. 3. ISRIC-World Soil Information, Wageningen, The Netherlands.

Luoto, M. & Hjort, J., 2005. Evaluation of current statistical approaches for predictive geomorphological mapping. *Geomorphology*. 67, 299–315.

Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, *154*(1), 138-152.

Martin, M. Z., Labbé, N., André, N., Wullschleger, S. D., Harris, R. D., & Ebinger, M. H. 2010. Novel multivariate analysis for soil carbon measurements using laser-induced breakdown spectroscopy. *Soil Science Society of America Journal*, *74*(1), 87-93.

Martin, M. P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., & Arrouays, D. 2011. Spatial distribution of soil organic carbon stocks in France. *Biogeosciences*, *8*(5), 1053-1065.

Mayr, T. R., Palmer, R. C., & Cooke, H. J. 2008. Digital soil mapping using legacy data in the Eden valley, UK. In *Digital Soil Mapping with Limited Data* Springer Netherlands. p. 291-301.

McBratney, A.B., M.L. Mendonça Santos, and B. Minasny. 2003. On digital soil mapping. *Geoderma*, 117, 3-52. doi:10.1016/S0016-7061(03)00223-4.

Mcivor, J. G., Ash, A. J., & Cook, G. D. 1995. Land condition in the tropical tallgrass pasture lands. 1. Effects on herbage production. *The Rangeland Journal*, *17*(1), 69-85.

MDA-Federal, 2004. Landsat GeoCover 1990/TM and ETM+2000 edition mosaics Tile N-

32-33. ETM-EarthSat-MrSID, 1.0, USGS, Sioux Falls, South Dakota.

Meersmans, J., Martin, M. P., Lacarce, E., De Baets, S., Jolivet, C., Boulonne, L., Lehmann, S., Saby, N.P.A., Bispo, A., & Arrouays, D. 2012. A high resolution map of French soil organic carbon. *Agronomy for Sustainable Development*, *32*(4), 841-851.

Meersmans, J., van Wesemael, B., De Ridder, F., Fallas Dotti, M., De Baets, S., & Van Molle, M. 2009. Changes in organic carbon distribution with depth in agricultural soils in northern Belgium, 1960–2006. *Global Change Biology*, *15*(11), 2739-2750.

Meersmans, J., van Wesemael, B., Goidts, E., Van Molle, M., De Baets, S., & De Ridder, F. 2011. Spatial analysis of soil organic carbon evolution in Belgian croplands and grasslands, 1960–2006. *Global Change Biology*, *17*(1), 466-479.

Metting, F. B., Smith, J. L. and Amthor, J. S. 1999. Science needs and new technology for soil carbon sequestration. In: Rosenberg N.J., Izaurralde R.C., Malone E.L. (Eds.). Carbon sequestration in soils: science, monitoring and beyond, Battelle Press, Columbus, Ohio, pp. 1–34.

Miklos, M., Short, M. G., McBratney, A. B., & Minasny, B. 2010. Mapping and comparing the distribution of soil carbon under cropping and grazing management practices in Narrabri, North-west New South Wales. *Soil Research*, 48(3), 248-257.

Miller, B. A., Koszinski, S., Wehrhan, M., & Sommer, M. 2015. Impact of multi-scale predictor selection for modeling soil properties, *Geoderma*, 239–240, 97–106.

Minasny, B. & McBratney, A. B. 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy.*Chemometrics and Intelligent Laboratory Systems*, 94, 72-79.

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.

Mulugeta, L., Karltun, E and Olsson, M. 2005. Assessing soil physical and chemical property responses to deforestation and subsequent cultivation in smallholders farming system in Ethiopia. *Agriculture, Ecosystems and Environment*, 105,373-368.

Murty, D., Kirschbaum, M.U.F., McMurtie, R.E. & McGilvray, A. 2002. Does conversion of forest to agricultural land change soil carbon and nitrogen? A review of the literature. *Global Change Biology*, 8 , 105–123.

Nave, L. E., Swanston, C. W., Mishra, U., & Nadelhoffer, K. J. 2013. Afforestation effects on soil carbon storage in the United States: a synthesis. *Soil Science Society of*

America Journal, *77*(3), 1035-1047.

Nelson, D.W. & Sommers, L.E. 1996. Total carbon, organic carbon, and organic matter. In D.L. Sparks, A.L. Page, P.A. Helmke, R.H. Loeppert, P.N. Soltanpour, M.A. Tabatabai, C.T. Johnston, M.E. Sumner (Eds). "Methods of soil analysis. Part 3: Chemical methods".Soil Science Society of America, Inc., American Society of Agronomy, Inc.: Madison, USA. p. 961–1010.

Nishina, K., Ito, A., Beerling, D. J., Cadule, P., Ciais, P., Clark, D. B. & Yokohata, T. 2014. Quantifying uncertainties in soil carbon responses to changes in global mean temperature and precipitation. *Earth System Dynamics*, *5*(1), 197-209.

Obalum, S. E., Watanabe, Y., Igwe, C. A., Obi, M. E., & Wakatsuki, T. 2012. Carbon stock in the solum of some coarse-textured soils under secondary forest, grassland fallow, and bare footpath in the derived savanna of south-eastern Nigeria. *Soil Research*, *50*(2), 157-166.

Odeh, I. O., Leenaars, J., Hartemink, A., & Amapu, I. 2012. The challenges of collating legacy data for digital mapping of Nigerian soils. In: Minasny,B., Malone, B.P., McBratney,A. (eds). Digital Soil Assessments and Beyond. Taylor & Francis Group, London, p 453-458.

Poeplau, C., & Don, A. 2013. Sensitivity of soil organic carbon stocks and fractions to different land-use changes across Europe. *Geoderma*, *192*, 189-201.

Poeplau, C., Don, A., Vesterdal, L., Leifeld, J., Van Wesemael, B. A. S., Schumacher, J., & Gensior, A. 2011. Temporal dynamics of soil organic carbon after land-use change in the temperate zone–carbon response functions as a model approach. *Global Change Biology*, *17*(7), 2415-2427.

Post, W.M. & Kwon, K.C. 2000. Soil carbon sequestration and land-use change: processes and potential. *Global Change Biology*, 6,317–327.

Post, W. M., Emanuel, W. R., Zinke, P. J., & Stangenberger, A. G. 1982. Soil carbon pools and world life zones. *Nature*, 298, 156 - 159 doi: 10.1038/298156a0

Post, W. M., Izaurralde, R. C., Mann, L. K., & Bliss, N. 1998. Monitoring and verifying soil organic carbon sequestration. *Carbon Sequestration in Soils*, 41-66.

Powers, J. S., Corre, M. D., Twine, T. E., & Veldkamp, E. 2011. Geographic bias of field observations of soil carbon stocks with tropical land-use changes precludes spatial extrapolation. *Proceedings of the National Academy of Sciences*, *108*(15), 6318-6322.

Quinlan, J.R. 1993. Combining instance-based and model-based learning. In: Proceedings of the Tenth International Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA. pp. 236-243.

R Development Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raji, B. A., & Ogunwole, J. O. 2006. Potential of soil carbon sequestration under various land use in the sub-humid and semi-arid savanna of Nigeria, Lessons from long term experiments. *International Journal of Soil Science, 1*, 33-43.

Ravilious, C., Kapos, V., Osti, M., Bertzky, M., Bayliss, J. L., Dahiru, S., & Dickson, B. 2010. Carbon, biodiversity and ecosystem services: Exploring co-benefits. Nigeria: Preliminary Results. UNEP_WCMC: Cambridge, UK. 12p.

Reuter, H.I. & Nelson, A. 2009. Geomorphometry in ESRI Packages. In: T. Hengl and H.I. Reuter (Eds.). Geomorphometry: concepts, software, applications. *Development in Soil Science*, 33, 269-291.

Ritchie, J. C., McCarty, G. W., Venteris, E. R., & Kaspar, T. C. 2007. Soil and soil organic carbon redistribution on the landscape. *Geomorphology*, *89*(1), 163-171.

Rosenbloom, N. A., Harden, J. W., Neff, J. C., & Schimel, D. S. 2006. Geomorphic control of landscape carbon accumulation. *Journal of Geophysical Research: Biogeosciences (2005–2012), 111*(G1).

Roshetko, J. M., Lasco, R. D., & Angeles, M. S. D. 2007. Smallholder agroforestry systems for carbon storage. *Mitigation and Adaptation Strategies for Global Change*, *12*(2), 219-242.

Rusco, E., Jones, R. & Bidoglio, G. 2001. Organic matter in the soils of Europe: present status and future trends. European Soil Bureau, Soil and Waste Unit, Institute for Environment and Sustainability. JRC Ispra, Institute Joint Research Centre, European Commission, Italy.

Six, J., Callewaert, P., Lenders,S., De Gryze, S., Morris, S.J., Gregorich, E.G., Paul,E.A. & Paustian, K. 2002. Measuring and understanding carbon storage in afforested soils by physical fractionation. *Soil Science Society of America Journal*, 66:1981–1987.

Smith, P. 2012. Soils and climate change. *Current Opinion in Environmental Sustainability*, *4*(5), 539-544.

Sowunmi, F. A. & Akintola, J.O., 2010. Effect of climatic variability on maize production in Nigeria. *Research Journal of Environmental and Earth Sciences, 2*(1), 19-30.

Sreenivas, K., Sujatha, G., Sudhir, K., Kiran, D. V., Fyzee, M. A., Ravisankar, T., & Dadhwal, V. K. 2014. Spatial Assessment of Soil Organic Carbon Density Through Random Forests Based Imputation. *Journal of the Indian Society of Remote Sensing*, *42*(3), 577-587.

Su, Y,Z., Liu, W,J., Yang, R., Chang, X.X. 2009. Changes in soil aggregate, carbon, and nitrogen storages following the conversion of cropland to alfalfa forage land in the marginal oasis of northwest China. *Environmental Management*, 43, 1061–1070. doi:10.1007/s00267-009-9284-x

Suuster, E., Ritz, C., Roostalu, H., Kolli, R., & Astover, A. 2012. Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data. *European Journal of Soil Science*, *63*(3), 351-359.

UNEP (United Nations Environment Programme) 2007: Global Environment Outlook (GEO-4); Nairobi http://unep.org/geo/geo4/media/.

UNESCO-ASGA, 1963. Geological map of Africa, 1:5,000,000, Paris.

USGS, 2006. Shuttle Radar Topography Mission, 1 Arc Second scenes. In: Global Land Cover facility - University of Maryland (Ed.), Maryland, US.

Vågen, T. G., Winowiecki, L. A., Abegaz, A., & Hadgu, K. M. 2013. Landsat-based approaches for mapping of land degradation prevalence and soil functional properties in Ethiopia. *Remote Sensing of Environment*, 134, 266-275.

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

Verinumbe, I. 1991. Agroforestry development in northeastern Nigeria. *Forest Ecology and Management*, *45*(1), 309-317.

Wasige, J. E., Groen, T. A., Rwamukwaya, B. M., Tumwesigye, W., Smaling, E. M. A., & Jetten, V. 2014. Contemporary land use/land cover types determine soil organic carbon stocks in south-west Rwanda. *Nutrient Cycling in Agroecosystems*, *100*(1), 19-33.

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil*, *340* (1-2), 7-24.

Wiesmeier, M., Barthold, F., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., Schilling' B., Angst, G., Lützow, M. & Kögel-Knabner, I. 2014. Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany). *Geoderma Regional*, 1, 67-78.

Wiesmeier, M., Spörlein, P., Geuß, U., Hangen, E., Haug, S., Reischl, A., Schilling,B., Lützow, M., & Kögel-Knabner, I. 2012. Soil organic carbon stocks in southeast Germany (Bavaria) as affected by land use, soil type and sampling depth. *Global Change Biology*, *18*(7), 2233-2245.

Wilson, B.R., Growns, I. & Lemon, J. 2008. Land-use effects on soil properties on the north-western slopes of New South Wales: implications for soil condition assessment. *Australian Journal of Soil Research*, 46:359–367.

Wynn, J. G., Bird, M. I., Vellen, L., Grand-Clement, E., Carter, J., & Berry, S. L. 2006. Continental-scale measurement of the soil organic carbon pool with climatic, edaphic, and biotic controls. *Global Biogeochemical Cycles*, 20(1). doi:10.1029/2005GB002576.

Yang, Y., Fang, J., Ma, W., Smith, P., Mohammat, A., Wang, S., & Wang, W. 2010. Soil carbon stock and its changes in northern China's grasslands from 1980s to 2000s. *Global Change Biology*, *16*(11), 3036-3047.

Young, R., Wilson, B. R., McLeod, M., & Alston, C. 2005. Carbon storage in the soils and vegetation of contrasting land uses in northern New South Wales, Australia. *Soil Research*, *43*(1), 21-31.

.

Chapter 5.

Enhancing pedotransfer functions with environmental attributes to estimate selected soil properties in a data-scarce situation

Submitted: **Akpa, S. I.C.,** I.O.A. Odeh, T.F.A. Bishop & S.U. Sabastine, 2016. Enhancing pedotransfer functions with environmental attributes for estimating bulk density and effective cation exchange capacity in a data-scarce situation. *Soil Use and Management* (Revised)

Chapter 5 Enhancing pedotransfer functions in a data scarce situation

Abstract

Soil bulk density (BD) and effective cation exchange capacity (ECEC) are among the most important soil properties required for crop growth and environmental management. This study aimed to explore the combination of soil and environmental data in developing pedo-transfer functions (PTFs) for BD and ECEC. Multiple linear regression (MLR) and Random forest model (RFM) were employed in developing PTFs using three different datasets: soil data (PTF-1), environmental data (PTF-2) and the combination of soil and environmental data (PTF-3). In developing the PTFs, three depth increments were also considered: all depth, topsoil (< 0.40 m) and subsoil (> 0.40 m). Results showed that PTF-3 (R^2; 0.29 to 0.69) outperformed both PTF-1 (R^2; 0.11 to 0.18) and PTF-2 (R^2; 0.22 to 0.59) in BD estimation. However, for ECEC estimation, PTF-3 (R^2; 0.61 to 0.86) performed comparably as PTF-1 (R^2; 0.58 to 0.76) with both PTFs out-performing PTF-2 (R^2; 0.30 to 0.71). Also, grouping of data into different soil depth increments improves the estimation of BD with PTFs (especially PTF-2 and PTF-3) performing better at subsoils than topsoils. Generally, the most important predictors of BD are sand, silt, elevation, rainfall, temperature for estimation at topsoil while while EVI, elevation, temperature and clay are the most important BD predictors in the subsoil. Also, clay, sand, pH, rainfall and SOC are the most important predictors of ECEC in the topsoil while pH, sand, clay, temperature and rainfall are the most important predictors of ECEC in the subsoil. Findings are important for overcoming the challenges of building national soil databases for large scale modelling in most data-scarce countries, especially in the Sub-Saharan Africa (SSA).

Keywords:

Pedotransfer function, digital soil mapping, predictive modelling, Multiple linear regression, Random forest, Nigeria.

Chapter 5 Enhancing pedotransfer functions in a data scarce situation

5.1 Introduction

Soil bulk density (BD) and effective cation exchange capacity (ECEC) are among the most important soil properties required in national soil databases (Arrouays et al., 2014). This can be attributed to their relevance in decision making on crop growth, soil use and environmental management. BD influences water and solute movement through the soil. It is vital for the estimation of soil carbon stocks and soil hydraulic properties which are important input parameters of process-based models for simulating the flux of water, nutrients and greenhouse gases (Bellamy et al., 2005; Ungaro et al., 2010; Ghehi et al., 2012). ECEC on the other hand is relevant in many soil, crop and environmental risk assessment models (Liao et al., 2014). It is a measure of the fertility, nutrient retention and pH buffering capacity of soils as well as the capacity of soils to protect groundwater from cation contamination (Noble et al., 2000; Akbarzadeh et al., 2009).

Despite their importance, adequate information on these two soil properties is usually lacking in the national soil databases of most data-sparse countries especially in Sub-Saharan Africa (SSA). This is owing to the fact that direct determination of these soil attributes is cumbersome and prohibitive, especially when required over relatively large land areas. Pedotransfer functions (PTFs), defined as predictive functions of certain soil properties derived from other easily measured properties (Bouma, 1989), has proven to be a useful and quick way of estimating scarce but very important soil properties from more easily obtainable soil data.

Several studies have been conducted to develop PTFs for predicting soil properties from basic soil data using various modelling techniques (Suuster et al., 2011; Ghehi et al., 2012; Haghverdi et al., 2012; Sequeira et al., 2014). Nonetheless, the reliability of many of these PTFs is largely dependent on the amount (data size) and structure (range) of the input parameters (Romano and Chirico, 2004; Ghehi et al., 2012; Haghverdi et al., 2012). For instance, in a relatively small area, with fairly homogeneous soil properties and topography, high reliability could be obtained from a reasonably few number of soil samples (Ghehi et al., 2012). However, in large and heterogeneous landscape with high spatial soil variability, reliability of PTFs is greatly impacted by the size and spread of soil sampling points. Considering this at the national scale, developing PTFs will require extensive soil sampling to capture the spatial variability of the target soil properties. This could be a herculean task as it involves huge investment in terms of man power and finance.

Thus, to improve the performance of PTFs, several techniques have been employed prior to fitting PTFs including, grouping of input data based on soil taxonomy (Heuscher et al., 2005) and the incorporation of soil physiographic and morphological attributes such as consistence and structure (Calhoun et al., 2001) and horizon designation (Jalabert et al., 2010). However, one major challenge is that most of these soil morphological attributes are not always available in soil survey reports. Other studies have reported improvement of PTFs following the combination of soil data (e.g sand, silt, clay, SOC) with environmental data such as percent slope, annual rainfall amount and vegetation indices (Leij et al. 2004; Sharma et al., 2006; Jana and Mohanty, 2011; Wang et al., 2014). For instance, Leij et al. (2004) used a combination of aspect, elevation, slope, potential solar radiation) and soil data (e.g. sand, silt, SOC) to develop PTFs for predicting soil hydraulic properties of soils in Basilicata, Italy. Their results showed significant improvement over the use of only soil data as inputs. Similarly, Sharma et al. (2006) reported that the combination of topographic indices (elevation, slope, aspect, and flow accumulation), vegetation indices (NDVI) and soil data (sand, clay, silt, SOC) in PTFs is a more reliable approach to estimating soil moisture contents. Recently, Wang et al. (2014) used the combination of clay, SOC, slope gradient and altitude to develop PTFs for estimating BD across the Loess Plateau in China. Although these studies have demonstrated improvement in the performance of PTFs using the combination of soil and environmental data over the use of only soil data, many of these studies were carried out at the field scale and mainly for moisture content prediction. It would be pertinent therefore to investigate the inclusion of environmental data in PTFs to predict other very important soil properties such as BD and ECEC, especially at the national level, which more often than not, is characterised by scarce soil data that are unevenly distributed. Hence, the objectives of this study are to (i) examine the performance of PTFs developed by combining soil and environmental data for the estimation of BD and ECEC at a national scale and (ii) examine the impact of data groupings on the performance of PTFs using MLR and RFM.

5.2 Materials and Methods

5.2.1 Description of the study area and soil profile distribution

This study was carried out using legacy soil data for Nigeria. Nigeria is located within latitudes 4° and 14° North, and longitudes 2° and 15° East, with a total area of about 923,768 km^2 (356,667 sq mi). The climate is tropical with two marked climatic contrasts: humid in the south (annual precipitation of 1250 to 4000 mm) and semi-arid in the north (annual precipitation of 500 to 1250 mm). According to the FAO soil taxonomy, major soils in Nigeria include Lixisols (24%), Arenosols (17%), Leptosols (13%), Fluvisols (7%), Plinthosols (5.5%), Luvisols (5%), Acrisols (4.5%), Gleysols (3.4%), Vertisols (1.6%), Ferralsols (1.5%) and Regosols (1.5 %). The spatial distribution of the soil profiles used in this study is shown in Fig. 5.1. Data on measured BD (for at least 3 soil layers) were available for 260 soil profiles and 1161 soil layers. Also, data on ECEC were available from 627 profiles comprising of 2124 layers. BD data are sparsely distributed across the country with denser distribution around the central, western and eastern regions. However, there is a good representation of the major soils in the BD data (Fig. 5.2a). ECEC data on the other hand is somewhat evenly distributed across the country with greater representation of major soils compared to BD data (Fig. 5.2b).

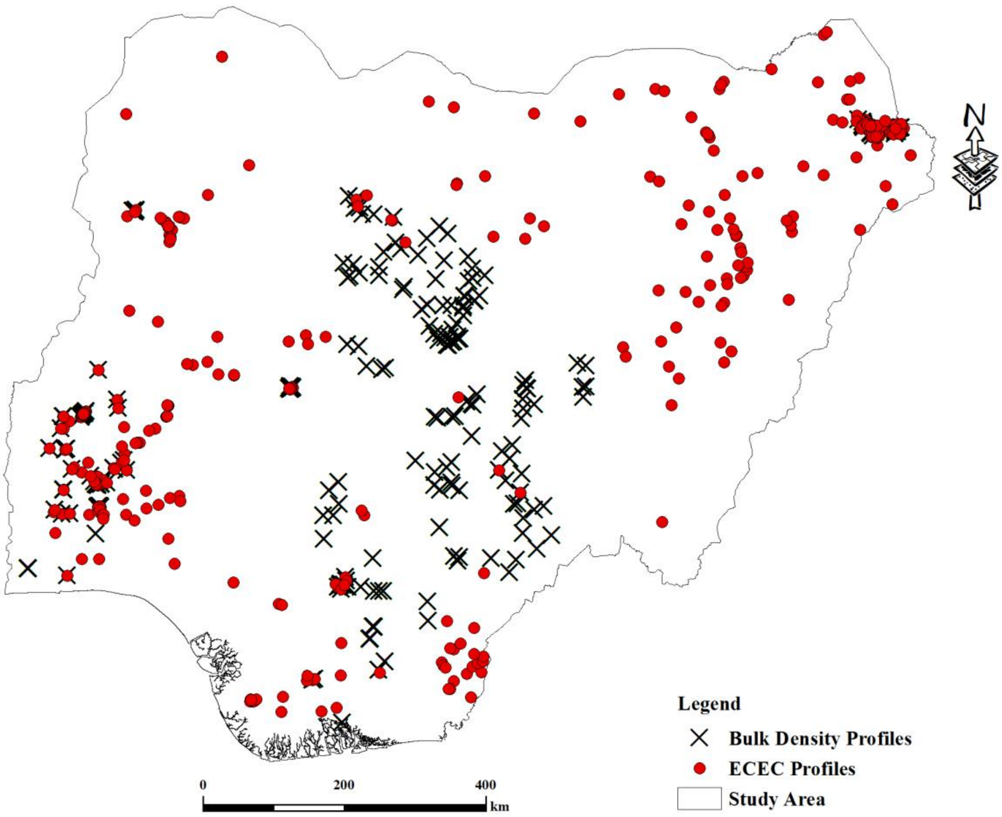
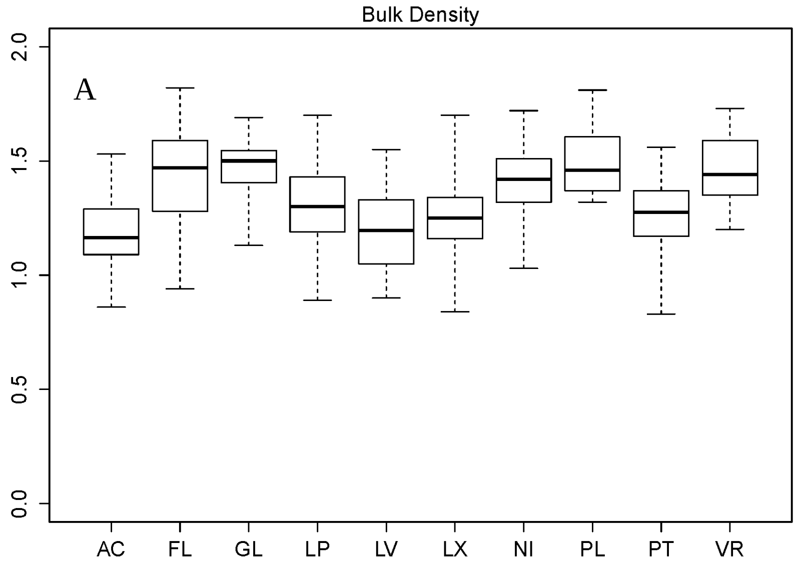


Figure 5.1. Spatial distribution of legacy soil profiles for Bulk density and Effective cation exchange capacity in Nigeria.

5.2.2 Predictor variables

5.2.2.1 Soil data

In this study, soil data such as soil organic carbon (SOC), pH, particle size fractions (sand, silt, clay) were selected as potential predictors of BD and ECEC. Soil data for Nigeria was extracted from the ISRIC African soil profile database version 1.0 (Leenaars, 2012) which were compiled from an array of soil surveys in Nigeria between 1960 and 2010. About 1141 soil profiles covering most part of the country were described and analysed in the database. In obtaining data for the database, ECEC was determined using cation summation of all the measured cations while BD was determined by core sampling and subsequent oven-drying. For details on laboratory methods and procedures in obtaining the soil data used in this study, readers are referred to Leenaars (2012) and Akpa et al. (2014; 2016).



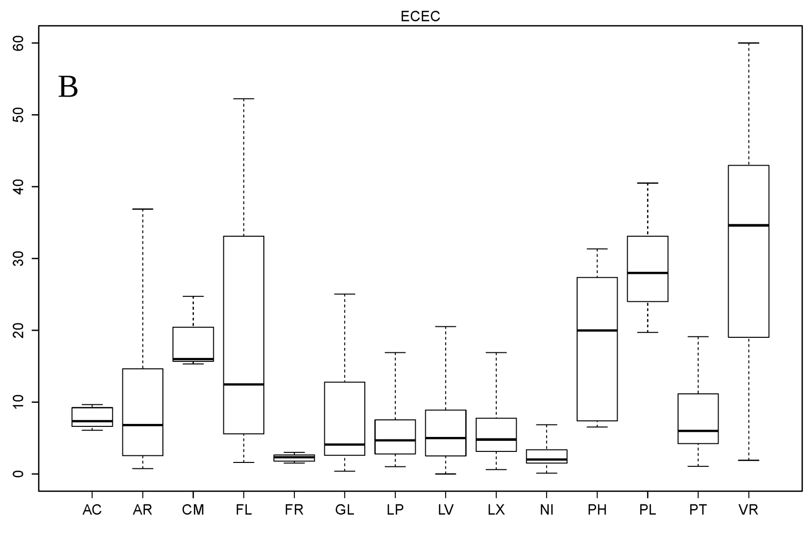


Figure 5.2. Boxplot of Bulk density (Mg m^{-3}) (A) and Effective cation exchange capacity (cmol kg^{-1}) (B) based on soils of Nigeria. Abbreviations: AC, Acrisols; AR, Arenosols; CM, Cambisols; FL, Fluvisols; FR, Ferralsols; GL, Gleysols; LP, Leptosols; LV, Luvisols; LX, Lixisols; NI, Nitisols; PH, Phaeozems; PL, Planosols; PT, Plinthosols; VR, Vertisols.

5.2.2.2 Environmental data

Environmental data selected as potential predictors of BD and ECEC in this study include topographic indices such as elevation, slope gradient, aspect, profile and plan curvatures, flow accumulation, topographic wetness index (TWI), stream power index (SPI). These were derived from SRTM 3 arc (90 m) digital elevation model (DEM) following Reuter and Nelson (2009). For further details on the topographic indices, see Moore et al. (1993) and Wilson and Gallant (2000). Other environmental data include 30 arc-second (1km) global mean annual rainfall and mean annual temperature data obtained from WorldClim (Hijmans, et al., 2005) as well as MODIS enhance vegetation index (EVI) and normalized vegetation index (NDVI) data. The summary statistics of the soil and environmental data used in developing PTFs for BD and ECEC in this study are shown in Table 5.1 and Table 5.2.

5.2.3 Data combinations and groupings

In this study, three different combinations of the various predictors mentioned in Section 5.2.2 above were used in developing PTFs for BD and ECEC: soil data (PTF-1), environmental data (PTF-2) and combination of soil and environmental data (PTF-3). Prior to deriving the PTF for BD and ECEC, we first grouped the entire dataset into three 3 different categories based on soil sampling depth: all data, topsoil (< 0.40 m) and subsoil (> 0.40 m) data. These groupings were based on the definition reported by Ghehi et al. (2012). Topsoil data have lower limits of their horizon layer ≤30cm, while subsoil data have upper limit of their horizon layer >30cm. However, when there is an overlap due to wide range of depth intervals, the separation between topsoil and subsoil was set at an upper limit of 40cm (Brahim et al. 2012). One obvious reason for the stratification of data by sampling depth is to account for the effect of soil management and anthropological activities on soil properties of surface and subsurface samples (Benites et al., 2007).

Table 5.1. Summary statistics of selected predictors of Bulk density.

	All data				Topsoil				Subsoil			
Attribute	max	min	mean	SD	max	min	mean	SD	max	min	mean	SD
BD (Mg m^{-3})	2.03	0.73	1.32	0.19	1.84	0.73	1.31	0.16	2.03	0.77	1.32	0.20
Sand (%)	100.00	6.20	55.22	22.95	96.00	8.40	61.10	21.90	100.00	6.20	52.12	22.9
Silt (%)	80.00	0.00	18.06	13.32	80.00	0.00	18.80	15.31	75.40	0.00	17.68	12.13
Clay (%)	84.00	0.00	26.72	16.78	69.00	0.00	20.10	14.67	84.00	0.00	30.21	16.78
TWI	13.40	3.01	6.16	2.00	13.40	3.01	6.02	1.94	13.40	3.01	6.24	2.03
Temp (^{o} K)	282.61	235.1	264.54	9.71	282.61	235.10	263.81	9.53	282.61	235.10	264.92	9.78
SPI	3.99	-8.42	-3.03	1.85	3.99	-8.42	-3.05	1.81	3.99	-8.42	-3.02	1.87
Slope (%)	6.97	0.01	0.88	0.81	6.97	0.01	0.92	0.81	6.97	0.01	0.86	0.8
ProfC	0.02	-0.02	$3.21x10^{-4}$	$3.66x10^{-3}$	0.02	-0.02	$3.00x10^{-4}$	$4.21x10^{-3}$	0.02	-0.02	$3.60x10^{-4}$	$3.34x10^{-3}$
PlanC	0.02	-0.02	$5.37x10^{-4}$	$3.92x10^{-3}$	0.02	-0.02	$5.30x10^{-4}$	$4.14x10^{-3}$	0.02	-0.02	$3.82x10^{-4}$	$3.79x10^{-3}$
Rainfall (mm)	2377.98	445.52	1237.71	328.42	2377.98	445.52	1274.68	333.6	2276.24	445.52	1218.24	324.21
FLACC	5978.00	0.00	59.33	427.89	5978.00	0.00	60.72	457.03	5978.00	0.00	58.60	412.05
EVI	0.50	0.15	0.33	0.07	0.50	0.15	0.33	0.07	0.5	0.15	0.33	0.07
Elevation (m)	931.16	15.12	331.46	243.50	931.16	15.12	339.19	251.93	931.16	19.77	327.39	239.02
Aspect (^{o})	353.60	5.71	196.29	92.16	353.60	5.71	195.53	93.17	353.6	5.71	196.69	91.68
NDVI	0.66	0.21	0.46	0.08	0.66	0.21	0.46	0.08	0.66	0.21	0.46	0.08

†BD, Bulk density; EVI, Enhanced vegetation index; FLACC, flow accumulation; Max; maximun; Min; minimum; NDVI, Normalized difference vegetation index; ProfC, profile curvature; PlanC, plan curvature; Temp, Temperature; TWI, Topographic wetness index; SD; standard deviation; SPI, Stream power index.

Table 5.2. Summary statistics of selected predictors of Effective cation exchange capacity.

	All Data				Topsoil				Subsoil			
Attribute	max	min	mean	SD	max	min	mean	SD	max	min	mean	SD
ECEC (cmolc kg^{-1})	60.00	0.00	10.07	11.34	60.00	0.75	11.97	13.53	54.00	0.00	9.05	9.83
pH	10.10	4.10	6.43	1.11	9.30	4.10	6.40	0.96	10.10	4.20	6.44	1.18
Sand (%)	100.00	0.20	54.67	24.07	100.00	0.40	56.28	26.77	100.00	0.20	53.81	22.46
Clay (%)	88.10	0.00	26.59	19.20	88.10	0.00	23.17	20.89	87.50	0.00	28.43	17.98
Temp (^{o} K)	286.03	251.42	269.66	6.70	286.03	251.42	270.30	7.06	286.03	251.42	269.32	6.48
Rainfall (mm)	2692.18	286.06	1172.63	559.14	2692.18	286.06	1172.38	630.08	2692.18	286.06	1172.76	517.48
Silt (%)	80.00	0.00	18.73	13.09	80.00	0.00	20.46	14.81	75.40	0.00	17.79	11.96
Elevation (m)	1098.27	7.78	273.79	156.55	1098.27	7.78	270.81	153.55	1098.27	7.78	275.38	158.2
NDVI	0.67	0.08	0.44	0.12	0.67	0.08	0.43	0.13	0.67	0.08	0.45	0.12
SOC (g kg^{-1})	111.00	0.00	5.48	7.52	111.00	0.20	10.18	10.44	49.6	0.00	2.97	3.22
EVI	0.53	0.06	0.33	0.10	0.53	0.06	0.32	0.10	0.51	0.06	0.33	0.09
Slope (%)	14.03	0.00	1.03	1.66	14.03	0.00	0.95	1.63	14.03	0.02	1.07	1.67
Aspect (^{o})	359.64	0.48	184.08	106.58	359.64	0.48	175.33	108.24	359.64	1.94	188.76	105.45
SPI	3.30	-10.72	-3.26	2.13	3.30	-10.72	-3.61	2.21	3.30	-7.69	-3.07	2.06
TWI	14.37	2.59	6.28	2.02	14.37	2.59	6.35	2.11	13.87	2.59	6.24	1.97
ProfC	0.03	-0.02	$4.16x10^{-4}$	$4.27x10^{-3}$	0.03	-0.02	$3.72x10^{-4}$	$4.28x10^{-3}$	0.03	-0.02	$4.39x10^{-4}$	$4.27x10^{-3}$
PlanC	0.05	-0.01	$4.84x10^{-4}$	$5.43x10^{-3}$	0.05	-0.01	$5.43x10^{-4}$	$4.93x10^{-3}$	0.05	-0.01	$4.52x10^{-4}$	$5.68x10^{-3}$
FLACC	4282.00	0.00	58.23	350.14	4282.00	0.00	52.69	340.89	4282.00	0.00	61.20	355.15

†ECEC, Effective cation exchange capacity; EVI, Enhanced vegetation index; FLACC, flow accumulation; Max; maximun; Min; minimum; NDVI, Normalized difference vegetation index; ProfC, profile curvature; PlanC, plan curvature; Temp, Temperature; TWI, Topographic wetness index; SD; standard deviation; SPI, Stream power index.

5.2.4 Deriving the Pedotransfer functions

Pedotransfer functions were derived using two prediction techniques: multiple linear regression (MLR) and Random forest model (RFM).

5.2.4.1 Multiple Linear Regression

MLR is a regression technique that explores a possible functional linear relationship between a primary variable and explanatory variables that are easy to measure (Burrough and McDonnell, 1998). MLR assumes that the data are independent of each other, normally distributed and homogeneous in variance. MLR is the most commonly used prediction techniques in PTF studies (Hollis et al., 2012) owing to its simplicity and ease of reproducibility. We employed MLR in this study through a stepwise linear regression (SLR) to develop PTFs of BD and ECEC using the aforementioned data combination and groupings. SLR was fitted so as to reduce the interference of redundant predictors with the performance of PTFs (Chan et al., 2011). Prior to fitting SLR model, preliminary checks in the form of descriptive statistics, scatter plots, correlation matrix, boxplot, residual plots etc. were carried out on the predictor variables to ensure they met standard criteria for fitting linear models.

5.2.4.2 Random Forest Model

RFM is a tree-based, robust prediction technique which was developed by Breiman (2001). RFM strength lies in its two randomization procedures of bootrapping and random input selection (Sequeira et al., 2014) and subsequent bagging of the predictions. Unlike MLR, the underlying tree of RFM can accommodate non-linear relationship between the response and predictor variables as well as interactions among the predictors (Akpa et al., 2016). Another major advantage of RFM over classical regression models such as MLR is in the determination of important predictors. While MLR uses stepwise and criterion-based procedures to select important predictors, RFM estimates the relative importance of the predictor variables, based on how worse the prediction would be if the data for a particular variable were permuted randomly (Prasad et al., 2006; Akpa et al., 2014). We implemented the RFM approach in R environment using the *randomForest* 4.6 package (R Core Team, 2014). Also, the 'importance" function in the random forest package was used to assess the relative importance of predictors in the various PTFs. Readers may contact the first author for a copy of the R script used in fitting RFM in this study.

5.2.5 Evaluation of the Pedotransfer functions.

The accuracy performance of the PTFs was tested by fitting the model on the calibration dataset and cross-validating the model using the validation dataset. To ensure stability and increase reliability of the PTFs, model calibration was done using 100 runs. At each run, the calibration and validation datasets were sampled at random using the ratio of 80:20 after which the performance of the models was evaluated using the difference between measured and estimated response variable. In doing this, three statistical indices based on mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2) and Lin's concordant correlation coefficient (P_c) were computed. The averages of the outputs from the 100 model runs were used as the final basis for comparison of different PTFs generated by the two prediction models. The following equations were used in computed the evaluation indices:

$$MAE = \frac{1}{N} \sum_{j=1}^{N} \left| M_j - P_j \right| \quad [5.1]$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \left(M_j - P_j \right)^2} \quad [5.2]$$

$$R^2 = \frac{\sum_{j=1}^{N}\left(P_j - M_j\right)^2}{\sum_{j=1}^{N}\left(M_j - \bar{M}_j\right)^2} \quad [5.3]$$

$$Pc = \frac{2p\sigma_m\sigma_p}{\sigma_m^2 + \sigma_p^2 + \left(\mu_m - \mu_p\right)^2} \quad [5.4]$$

where, M_j, $\bar{M}_j$ and P_j are the measured, mean of measured and predicted response variables, μ_o and μ_p are the means of the measured and predicted response variable while σ_m^2 and σ_p^2 are the corresponding variance and ρ the Pearson correlation coefficient between the raw and predicted variables. For good reliability or performance of the PTFs, MAE and RMSE should be as low as possible (i.e. close to zero) while Pc and R^2 should be high as possible (i.e. close to one).

5.3 Results

5.3.1 Relationship between predictor and target soil properties

The relationship between pairs of all potential predictors of BD and ECEC is shown in Table 5.3 and Table 5.4. There is a significant ($p < 0.05$) but relatively weak relationship between BD and most of the predictor variables, with the strongest relationship existing between BD and temperature (r = 0.35), elevation (r = 0.31) and sand (r = 0.28). This implies that these variables will be good candidates for predicting BD, especially in using the MLR model. In the case of ECEC, there is a stronger relationship with soil attributes and vegetation indices than topographic attributes. Clay content showed the strongest positive ($p < 0.01$) relationship (r = 0.57) with ECEC while sand content displayed the strongest negative correlation (r = -0.56) with ECEC. This is closely followed by pH (r = 0.49), EVI (r = - 0.49) and

NDVI (r = -0.46). In addition, SPI (r = -0.30) and rainfall (r = -0.40) showed a moderate relationship with ECEC.

5.3.2 Performance of PTFs

The summary of the model reliability indices for the BD PTFs developed based on the different data groupings is presented in Table 5.5 while the results of the ECEC PTFs are summarized in Table 5.6 respectively.

5.3.2.1 Bulk density

Generally, the prediction performances of PTFs used in estimating BD are appreciably low, especially with the use of MLR (Table 5.5). The RMSE values for MLR range from 0.14 to 0.19 Mg m^{-3} while the *Pc* ranged from 0.19 to 0.50 respectively. However grouping the model input data sets based on soil depth increment produced slightly better prediction performances at the subsoil (*Pc* of 0.23 to 0.50) compared to the topsoil (Pc of 0.26 to 0.43) or all profile data (Pc of 0.19 to 0.45). Considering the use of different combinations of predictors, the PTFs prediction performance and accuracy were better with the combination of soil environmental data (PTF-3) than either soil data (PTF-1) or environmental data (PTF-2). In all cases, PTF-1 gave higher prediction errors (RMSE; 0.149 to 0.182 Mg m^{-3}) than PTF-2 (RMSE; 0.11 to 0.13 Mg m^{-3}). Generally, RFM performed better than MLR with a RMSE in the range of 0.11 to 0.18 Mg m^{-3} and *Pc* values of 0.38 to 0.84.

Table 5.3. Pearson's correlation coefficient between Bulk density and predictors.

	BD	Sand	Silt	Clay	TWI	Temp	SPI	Slope	ProfC	PlanC	Rainfall	Flacc	EVI	Elevation	Aspect	NDVI	logsilt	logslope	logFLACC	logelev
BD	1.000																			
Sand	0.283	1.000																		
Silt	-0.182	-0.691	1.000																	
Clay	-0.242	-0.819	0.152	1.000																
TWI	0.052	-0.146	0.177	0.060	1.000															
Temp	0.347	0.191	-0.103	-0.179	0.190	1.000														
SPI	-0.215	0.164	-0.145	-0.109	0.350	-0.223	1.000													
Slope	-0.036	0.285	-0.298	-0.153	-0.482	-0.053	0.353	1.000												
ProfC	-0.017	-0.039	0.035	0.025	0.275	-0.021	0.274	-0.019	1.000											
PlanC	0.092	0.083	-0.107	-0.029	-0.576	-0.001	-0.383	0.292	-0.494	1.000										
Rainfall	0.020	0.154	-0.187	-0.062	-0.315	-0.217	0.230	0.265	-0.116	0.121	1.000									
FLACC	-0.013	0.040	-0.058	-0.009	0.412	0.036	0.383	-0.068	0.081	-0.095	0.009	1.000								
EVI	-0.008	0.216	-0.278	-0.076	-0.274	-0.020	0.349	0.365	-0.022	0.068	0.629	0.069	1.000							
Elevation	-0.313	-0.222	0.164	0.173	-0.096	-0.707	-0.009	-0.052	0.020	0.007	-0.302	-0.097	-0.335	1.000						
Aspect	-0.105	-0.019	0.035	-0.002	0.023	-0.053	0.057	-0.011	0.023	0.008	-0.057	0.080	-0.110	0.115	1.000					
NDVI	-0.124	0.195	-0.246	-0.071	-0.304	-0.142	0.393	0.395	0.012	0.041	0.606	0.086	0.940	-0.211	-0.052	1.000				
logsilt	-0.210	-0.360	0.591	0.023	0.132	-0.093	-0.030	-0.194	0.034	-0.127	-0.329	-0.014	-0.100	0.211	0.099	-0.033	1.000			
logslope	-0.228	0.271	-0.283	-0.146	-0.616	-0.360	0.522	0.735	-0.020	0.202	0.480	-0.053	0.543	0.080	0.027	0.607	-0.146	1.000		
logFLACC	-0.106	-0.029	0.027	0.019	0.578	-0.002	0.676	0.006	0.291	-0.508	-0.080	0.261	0.024	-0.027	-0.097	0.032	0.076	0.041	1.000	
logELevation	-0.334	-0.107	0.059	0.099	-0.100	-0.556	0.031	0.034	0.019	0.010	-0.432	-0.108	-0.322	0.909	0.125	-0.198	0.206	0.117	0.005	1.000

†BD, Bulk density;nEVI, Enhanced vegetation index; FLACC, flow accumulation; NDVI, Normalized difference vegetation index; ProfC, profile curvature; PlanC, plan curvature; Temp, Temperature; TWI, Topographic wetness index; SPI, Stream power index.

Table 5.4. Pearson's correlation coefficient between Effective cation exchange capacity and predictors.

	ECEC	pH	Sand	Clay	Temp	Rainfall	Silt	Elevation	NDVI	SOC	EVI	Slope	Aspect	SPI	TWI	ProfC	PlanC	Flacc
ECEC	1.000																	
pH	0.491	1.000																
Sand	-0.563	-0.032	1.000															
Clay	0.568	0.004	-0.841	1.000														
Temp	0.143	0.248	0.029	-0.009	1.000													
Rainfall	-0.396	-0.646	0.077	-0.034	-0.296	1.000												
Silt	0.202	0.051	-0.607	0.081	-0.041	-0.091	1.000											
Elevation	0.108	0.295	-0.054	0.008	0.003	-0.509	0.09	1.000										
NDVI	-0.461	-0.595	0.165	-0.078	-0.265	0.760	-0.19	-0.252	1.000									
SOC	0.104	-0.115	-0.123	0.117	-0.101	0.280	0.053	-0.123	0.167	1.000								
EVI	-0.491	-0.584	0.197	-0.113	-0.281	0.765	-0.199	-0.293	0.967	0.157	1.000							
Slope	-0.198	-0.206	0.095	-0.076	-0.174	0.188	-0.061	0.252	0.297	0.024	0.268	1.000						
Aspect	-0.107	-0.013	0.087	-0.061	0.075	-0.036	-0.071	-0.066	-0.028	-0.071	-0.007	-0.033	1.000					
SPI	-0.302	-0.278	0.162	-0.147	-0.312	0.276	-0.081	0.106	0.382	0.066	0.369	0.425	0.001	1.000				
TWI	0.243	0.171	-0.167	0.085	0.247	-0.224	0.183	-0.028	-0.311	0.006	-0.287	-0.417	0.052	0.144	1.000			
ProfC	-0.069	-0.116	0.048	-0.062	-0.088	0.237	0.001	-0.105	0.151	0.003	0.168	0.128	-0.167	0.168	0.053	1.000		
PlanC	-0.018	0.036	0.021	-0.028	-0.012	-0.046	0.004	0.173	0.058	-0.026	0.033	0.586	0.056	-0.141	-0.471	-0.085	1.000	
FLACC	-0.057	-0.034	0.034	-0.102	-0.016	-0.032	0.087	-0.001	-0.031	-0.006	-0.008	-0.057	-0.039	0.393	0.479	0.012	-0.173	1.000

†ECEC, Effective cation exchange capacity; EVI, Enhanced vegetation index; FLACC, flow accumulation; NDVI, Normalized difference vegetation index; ProfC, profile curvature; PlanC, plan curvature; Temp, Temperature; TWI, Topographic wetness index; SPI, Stream power index.

5.3.2.2 Effective Exchangeable cation exchange capacity

Table 5.6 presents the performance indices of ECEC PTFs using both MLR and RFM. Generally, the prediction performances of PTFs used in estimating ECEC are within medium to high range, especially with the use of RFM. MLR yielded RMSE values for the various PTFs in the range of 6.56 to 7.55 cmol kg^{-1} (PTF-1), 6.45 to 7.06 cmol kg^{-1} (PTF-3), and 8.79 to 10.54 cmol kg^{-1} (PTF-2), while RFM gave RMSE values in the range of 4.63 to 5.59 cmol kg^{-1} (PTF-1), 4.43 to 4.91 cmol kg^{-1} (PTF-2) and 5.69 to 8.38 cmol kg^{-1} (PTF-3), respectively. Comparing the *Pc* values of both prediction techniques, MLR derived PTFs for ECEC gave Pc values in the range of 0.72 to 0.81 (PTF-1), 0.73 to 0.82 (PTF-3), and 0.37 to 0.57 (PTF-2) while RFM produced Pc values in the range of 0.80 to 0.90 (PTF-1), 0.87 to 0.92 (PTF-3), and 0.75 to 0.83 (PTF-2).

Table 5.5. Model validation indices for predicting Bulk density.

	MLR				RFM			
Model	MAE	RMSE	R^2	Pc	MAE	RMSE	R^2	Pc
				All data				
PTF-1	0.140	0.179	0.109	0.185	0.135	0.176	0.182	0.380
PTF-2	0.132	0.167	0.220	0.366	0.092	0.122	0.588	0.751
PTF-3	0.122	0.161	0.294	0.451	0.077	0.107	0.689	0.800
				Topsoil				
PTF-1	0.116	0.152	0.161	0.255	0.114	0.149	0.207	0.398
PTF-2	0.118	0.149	0.183	0.306	0.100	0.133	0.348	0.501
PTF-3	0.108	0.141	0.280	0.431	0.087	0.118	0.489	0.608
				Subsoil				
PTF-1	0.151	0.189	0.139	0.23	0.140	0.182	0.221	0.421
PTF-2	0.138	0.174	0.250	0.404	0.081	0.110	0.703	0.822
PTF-3	0.128	0.166	0.327	0.496	0.073	0.102	0.754	0.839

† PTF-1, Pedotransfer functions using only soil data as predictors; PTF-2, Pedotransfer functions using only environmental data as predictors; PTF-3, Pedotransfer functions using the combination of soil and environmental data as predictors; MAE, mean absolute error; RMSE, root mean square error; R^2, coefficient of determination; Pc, Lin's concordance correlation coefficient.

5.3.3 Relative importance of predictors

The importance of each predictor in model performance is presented in Fig 5.3 and Fig. 5.4. The most important predictors of BD are sand, silt, elevation, rainfall, EVI, and temperature (Fig. 5.3). In the case of ECEC, the top five important predictors include pH, sand, clay, EVI and silt (Fig. 5.4). Together these attributes contributes to about 33% increase in prediction accuracy of ECEC PTFs (Fig. 5.4a). Dividing the input datasets into subsoil and topsoil significantly affected the relative important of these predictors. For example, sand and silt greatly influenced the prediction of BD at the topsoil (Fig. 5.3b) but their influence was masked by other attributes in the subsoil (Fig. 5.3c). Also, SOC and silt were very important predictors of ECEC at the topsoil (Fig. 5.4b) but their influence decreased with soil depth (Fig. 5.4c).

Table 5.6. Model validation indices for predicting Effective cation exchange capacity.

	MLR				RFM			
Model	MAE	RMSE	R^2	Pc	MAE	RMSE	R^2	Pc
				All data				
PTF-1	5.278	7.420	0.577	0.730	3.521	5.523	0.762	0.857
PTF-2	6.916	9.513	0.298	0.459	3.707	6.125	0.708	0.823
PTF-3	5.081	7.011	0.611	0.759	2.728	4.445	0.853	0.909
				Topsoil				
PTF-1	5.393	7.532	0.691	0.812	3.563	5.446	0.835	0.907
PTF-2	7.909	10.543	0.388	0.559	5.679	8.388	0.622	0.751
PTF-3	5.346	7.397	0.702	0.821	3.303	5.134	0.860	0.913
				Subsoil				
PTF-1	4.487	6.437	0.566	0.721	3.085	4.627	0.801	0.879
PTF-2	6.177	8.572	0.229	0.373	3.300	5.594	0.680	0.798
PTF-3	4.494	6.394	0.583	0.732	2.712	4.469	0.805	0.870

†PTF-1, Pedotransfer functions using only soil data as predictors; PTF-2, Pedotransfer functions using only environmental data as predictors; PTF-3, Pedotransfer functions using the combination of soil and environmental data as predictors; MAE, mean absolute error; RMSE, root mean square error; R^2, coefficient of determination; Pc, Lin's concordance correlation coefficient.

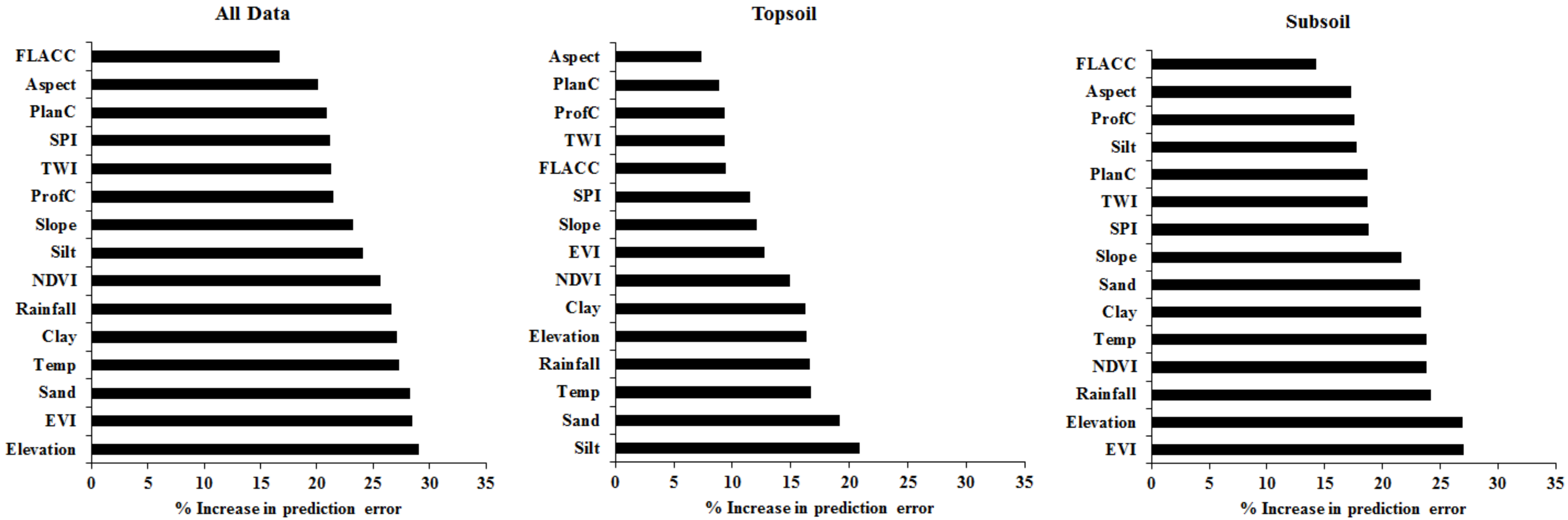


Figure 5.3. Importance of predictor variables PTF of BD as derived by Random Forest based on three sampling depths. Abbreviations: BD, Bulk density;nEVI, Enhanced vegetation index; FLACC, flow accumulation; NDVI, Normalized difference vegetation index; ProfC, profile curvature; PlanC, plan curvature; Temp, Temperature; TWI, Topographic wetness index; SPI, Stream power index

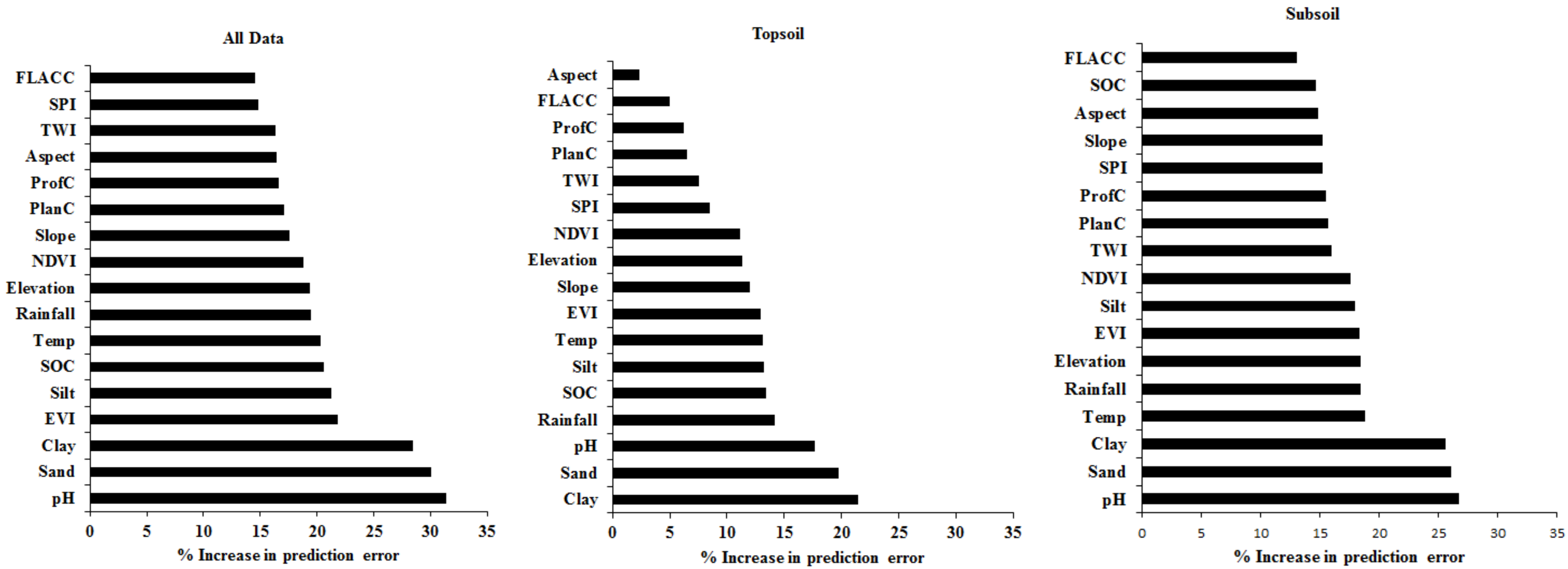


Figure 5.4. Importance of predictor variables for PTF of Effective Cation Exchange Capacity (ECEC) as derived by Random Forest based on three sampling depths. Abbreviations: ECEC, Effective cation exchange capacity; EVI, Enhanced vegetation index; FLACC, flow accumulation; NDVI, Normalized difference vegetation index; ProfC, profile curvature; PlanC, plan curvature; Temp, Temperature; TWI, Topographic wetness index; SPI, Stream power index.

5.4 Discussion

5.4.1 Performance of PTF models

Our reported RMSE and R^2 values are within acceptable and reported ranges for soils with similar properties (Benites, et al., 2007; Ghehi et al., 2012; Botula et al., 2015). For instance, our RMSE for RFM PTFs is within the range reported by Ghehi et al. (2012) for BD PTFs in Rwanda using k-NN but it is slightly better than the performance of PTFs derived using BRT. However, our MLR PTFs performed slightly poorer than those reported for highly weathered soils in Central Africa (Botula et al., 2015). The difference between our result and those of Botula et al. (2015) could be attributed to the difference in the datasets used, the varying extent of both studies as well as the weaker correlation between BD and predictors used in this study (Table 5.3) compared to theirs. There is more variation in our dataset since we have used legacy data which is a combination of data from disparate sources. In addition Botula et al. (2015) used a combination of five (5) soil physico-chemical data for BD PTF while we have used particle size fraction (sand, silt, clay) in this study. The superiority of RFM over MLR (Tables 5.5 and Table 5.6) could be attributed to the fact that RFM being non-parametric was able to capture, the complex relationships existing between the multivariable predictors used in this study. According to Merdun (2010), the relationship between soil properties and environmental attributes is complex and nonlinear. Notwithstanding the superiority of RFM PTFs, MLR derived PTFs are easier to reproduce as shown by the equations in Appendix 5.1 and Appendix 5.2.

5.4.2 Data groupings and reliability of PTFs

Our results show that irrespective of the prediction technique used, PTFs performed better on subsoil than the topsoil data (Table 5.4 and Table 5.5), especially with BD PTFs. This is in contrast to previous reports that grouping model input data by soil depth does not improve the prediction of BD in tropical soils (Botula et al., 2015 and Vos et al., 2005). The differences in our results may be due to the differences in the range of the datasets used and the level of disturbance of the topsoils in the study areas and (Hollis et al., 2012). Surface horizons undergo significant changes over time due to disturbances emanating from land use and trafficking. Therefore, physical soil properties such as BD are more stable in the subsoil than in the topsoil.

5.4.3 Inclusion of environmental data and the performance of PTFs

Our study shows a better performance of BD PTFs using combination of soil and environmental data compared to either soil or environmental data. This is consistent with the report of Wang et al. (2014). However, the performance of PTFs developed using RFM in this study is slightly better than those reported by Wang et al. (2014) using ANN. In contrast, MLR PTFs in their study is better. This could be attributed largely to differences in the sources as well as the density of data used in both studies. While we used low dense legacy BD data with its inherent limitations, they have sampled their BD data. This is further evident in the relative higher relationship between BD and other soil data in their study compared to ours (Table 5.3). Our study also reveals a superior performance of PTF-2 over PTF-1 in predicting BD. This suggests that environmental attributes are good alternatives to particle size fractions (clay, silt and sand) in BD prediction, especially in data-scarce situation. However, it is noteworthy that since we have used only 3 soil data (clay, silt and sands content), the poor performance of PTF-1 should not be overgeneralized as the addition of other soil data with higher explanatory power such as SOC and dithionite–citrate–bicarbonate-extractable Fe (DCB-Fe) and Al (DCB-Al) could significantly improve PTF-1. For ECEC, PTF-1 performs comparably with PTF-3 but slightly better than PTF-2. This suggests the adequacy of soil data in predicting soil ECEC in our study area.

5.4.4 Importance of predictor variables

Our result showed that important predictors of BD are sand, silt, elevation, rainfall, EVI, and temperature (Fig. 5.3) which is consistent with previous reports (Akpa et al., 2016). Elevation, in combination with climatic variables, influences the redistribution of particles during pedogenic processes (Adhikari et al., 2014; Akpa et al., 2014) and thereby could better explain variability in the BD of soils. Also, pH, sand, clay, SOC, silt, rainfall, temperature and elevation were identified as important predictors for ECEC (Fig. 4). This corroborates previous reports that more than 50% of the variation of cation exchange capacity could be explained by the variation in clay content, SOC as well as pH (Bell and Keulen, 1995; Krogh et al., 2000). Our results also showed that dividing the input datasets based on soil layer can affect the relative important of the predictor variables (Figs. 5.3 and 5.4). This is consistent with the report of Suuster et al. (2011) that the importance of predictor variables will vary based on the prediction condition and modelling criteria. The lesser influence of SOC on the variation of ECEC at the subsoil relative to topsoil (Fig. 5.4) has been previously reported (Asadu and Akamigbo, 1990) and could be attributed to the low accumulation of soil organic matter in the subsoil compared to topsoil.

5.5 Conclusions

This study demonstrates an effort at predicting soil attributes, over a relatively large extent under data-sparse situation using a combination of soil and environmental data as predictors. Across all data groupings, the combination of soil and environmental data gave higher prediction accuracy for BD. However, the incorporation of environmental data show no significant improvement in the prediction of ECEC over the use of only soil attributes. Generally, the use of soil data gave a better prediction of ECEC than the use of only environmental data. This study showed that in Nigeria, sand, silt, temperature, rainfall and elevation, NDVI and EVI are the most important predictors of BD at the topsoil while EVI, elevation, temperature and clay are the most important BD predictors in the subsoil. Also, clay, sand, pH, rainfall and SOC are the most important predictors of ECEC in the topsoil while pH, sand, clay, temperature and rainfall are the most important predictors of ECEC in the subsoil. Findings from this study are important in overcoming the daunting challenges of building national soil databases for large scale modelling in most data-scarce countries, especially in the Sub-Saharan Africa (SSA).

References

Adhikari, K., Hartemink, A.E., Minasny, B., Kheir, R.B., Greve, M.B & Greve, M.H. 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. *PLoS One*, 9 (8), e105519.

Agyare, W. A. 2004. Soil characterization and modeling of spatial distribution of saturated hydraulic conductivity at two sites in the Volta Basin of Ghana (No. 17). Cuvillier Verlag. 194 p.

Akbarzadeh, A., TaghizadehMehrjardi, R., Rahimi Lake, H., & Ramezanpour, H. 2009. Application of artificial intelligence in modeling of soil properties (case study: Roodbar region, north of Iran). *Environmental. Research. Journal*, *3*(2), 19-24.

Akpa, S. I.C, Odeh, I. O. A, Bishop, T. F.A, & Hartemink, A. E. 2014. Digital mapping of soil particle-size fractions for Nigeria. *Soil Science Society of America Journal*, *78*(6), 1953-1966.

Akpa, S. I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E. & Amapu, I.Y. 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma*, 271, 202-215.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonça Santos, M.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A. & Zhang, G.L. 2014. Chapter Three-GlobalSoilMap: toward a fine-resolution global grid of soil properties. *Advances in Agronomy*. 125, 93–134.

Asadu, C.L.A. & Akamigbo, F.O.R. 1990. Relative contribution of organic matter and clay fractions to catio exchange capacity of soils in southern Nigeria. *Samaru Journal of Agricultural Research*, 7:17-23.

Bellamy, P. H., Loveland, P. J., Bradley, R. I., Lark, R. M., & Kirk, G. J. 2005. Carbon losses from all soils across England and Wales 1978–2003. *Nature*, *437*(7056), 245-248.

Bernoux, M., Cerri, C., Arrouays, D., Jolivet, C., & Volkoff, B. 1998. Bulk densities of Brazilian Amazon soils related to other soil properties. *Soil Science Society of*

America Journal, *62*(3), 743-749.

Benites, V. M., Machado, P. L., Fidalgo, E. C., Coelho, M. R., & Madari, B. E. 2007. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. *Geoderma*, *139*(1), 90-97.

Blake, G.R. and K.H Hartge 1986. Bulk density. In: Klute, A. (ed). Methods of soil analysis, part 1. Physical and mineralogical methods. Am. Soc. Agron. Madison, USA. p 365 – 375.

Botula, Y. D., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J., & Cornelis, W. M. 2015. Hierarchical Pedotransfer Functions to Predict Bulk Density of Highly Weathered Soils in Central Africa. *Soil Science Society of America Journal*, *79*(2), 476-486.

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. *Advances in Soil Science*, 9:177-213.

Brahim, N., Bernoux, M., & Gallali, T. 2012. Pedotransfer functions to estimate soil bulk density for Northern Africa: Tunisia case. *Journal of Arid Environments*, *81*, 77-83.

Breiman, L. 2001. Random forests. *Machine learning*, *45*(1), 5-32.

Burrough, P.A., & McDonnell, R.A. 1998. Principles of Geographical Information Systems. Oxford University Press, Oxford, p. 333.

Calhoun, F. G., Smeck, N. E., Slater, B. L., Bigham, J. M., & Hall, G. F. 2001. Predicting bulk density of Ohio soils from morphology, genetic principles, and laboratory characterization data. *Soil Science Society of America Journal*, *65*(3), 811-819.

Chan, K. Y., Kwong, C. K., Dillon, T. S., & Tsim, Y. C. 2011. Reducing overfitting in manufacturing process modeling using a backward elimination based genetic programming. *Applied Software Computing*, *11*(2), 1648-1656.

Chirico, G. B., Medina, H., & Romano, N. 2010. Functional evaluation of PTF prediction uncertainty: An application at hillslope scale. *Geoderma*, *155*(3), 193-202.

De Vos, B., Van Meirvenne, M., Quataert, P. Deckers, J. Muys, B. 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. *Soil Science Society of America Journal*, 69: 500-510.

Ghehi, N. G., Nemes, A., Verdoodt, A., Van Ranst, E., Cornelis, W. M., & Boeckx, P. 2012. Nonparametric techniques for predicting soil bulk density of tropical rainforest topsoils in Rwanda. *Soil Science Society of America Journal*, *76* (4), 1172-1183.

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*, 46-54.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. The Elements of Statistical Learning. 2nd edition. Springer, Berlin 739p.

Heuscher, S. A., Brandt, C. C., & Jardine, P. M. 2005. Using soil physical and chemical properties to estimate bulk density. *Soil Science Society of America Journal*, *69*(1), 51-56.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*. 25, 1965–1978.

Hollis, J. M., Hannam, J., & Bellamy, P. H. 2012. Empirically-derived pedotransfer functions for predicting bulk density in European soils. *European Journal of Soil Science*, *63*(1), 96-109.

Jalabert, S. S. M., Martin, M. P., Renaud, J. P., Boulonne, L., Jolivet, C., Montanarella, L., & Arrouays, D. 2010. Estimating forest soil bulk density using boosted regression modelling. *Soil Use and Management*, *26*(4), 516-528.

Jana, R. B., & Mohanty, B. P. 2011. Enhancing PTFs with remotely sensed data for multi-scale soil water retention estimation. *Journal of hydrology*, *399*(3), 201-211.

Kätterer, T., Andrén, O., & Jansson, P. E. 2006. Pedotransfer functions for estimating plant available water and bulk density in Swedish agricultural soils. *Acta Agriculturae Scandinavica Section B-Soil and Plant Science*, *56*(4), 263-276.

Keller, T., & Håkansson, I. 2010. Estimation of reference bulk density from soil particle size distribution and soil organic matter content. *Geoderma*, *154*(3), 398-406.

Leenaars, J. G. B. 2012. Africa soil profiles database, Version 1.0. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). ISRIC Rep, 3.

Liao, K., Xu, S., Wu, J., Zhu, Q., & An, L. 2014. Using support vector machines to predict cation exchange capacity of different soil horizons in Qingdao City, China. *Journal of Plant Nutrition and Soil Science*, *177*(5), 775-782.

Leij, F.J., N. Romano, M. Palladino, and M.G. Schaap. 2004. Topographical attributes to predict soil hydraulic properties along a hillslope transect. *Water Resources Research* 40:1–15.

Merdun, H. 2010. Alternative methods in the development of pedotransfer functions for soil hydraulic characteristics. *Eurasian Soil Science*, *43*(1), 62-71.

Minasny, B., & Hartemink, A. E. 2011. Predicting soil properties in the tropics. *Earth-Science Reviews*, *106*(1), 52-62.

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

Noble, A. D., Gillman, G. P., & Ruaysoongnern, S. 2000. A cation exchange index for assessing degradation of acid soil by further acidification under permanent agriculture in the tropics. *European Journal of Soil Science*, *51*(2), 233-243.

Pachepsky, Ya. A., Timlin, D. J., & Rawls, W. J. 2001. Soil water retention as related to topographic variables. *Soil Science Society of America Journal*, 65, 1787–1795.

Prasad, A. M., Iverson, L.R., &Liaw, A. 2006Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9: 181-199.

R Development Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reuter, H.I. & Nelson, A. 2009. Geomorphometry in ESRI Packages. In: T. Hengl and H.I. Reuter (Eds.). Geomorphometry: concepts, software, applications. *Development in Soil Science 33*, 269-291.

Romano, N., & Chirico, G. B. 2004. The role of terrain analysis in using and developing pedotransfer functions. *Developments in Soil Science*, *30*, 273-294.

Sequeira, C. H., Wills, S. A., Seybold, C. A., & West, L. T. 2014. Predicting soil bulk density for incomplete databases. *Geoderma*, *213*, 64-73.

Seybold, C. A., Grossman, R. B., & Reinsch, T. G. 2005. Predicting cation exchange capacity for soil survey using linear models. *Soil Science Society of America Journal*, *69*(3), 856-863.

Sharma, S. K., Mohanty, B. P., & Zhu, J. 2006. Including topography and vegetation attributes for developing pedotransfer functions. *Soil Science Society of America Journal*, *70*(5), 1430-1440.

Suuster, E., Ritz, C., Roostalu, H., Reintam, E., Kõlli, R., & Astover, A. 2011. Soil bulk density pedotransfer functions of the humus horizon in arable soils. *Geoderma*, *163*(1), 74-82.

Ungaro, F., Staffilani, F., & Tarocco, P. 2010. Assessing and mapping topsoil organic carbon stock at regional scale: a scorpan kriging approach conditional on soil map delineations and land use. *Land Degradation & Development*, *21*(6), 565-581.

Wang, Y., Shao, M. A., Liu, Z., & Zhang, C. 2014. Prediction of Bulk Density of Soils in the Loess Plateau Region of China. *Surveys in Geophysics*, *35*(2), 395-413.

Wilson, J.P. & Gallant, J.C. 2000. Digital Terrain Analysis. In: Wilson JP and. GallantJC (eds) Terrain Analysis: Principles and Application. John Wiley and Sons. pp 1-27.

Chapter 6.

Multi-criteria evaluation of irrigation suitability at national scale: Application of Choquet Integral

Akpa, S. I.C., I.O.A. Odeh, T.F.A. Bishop & K.O. Oluwasemire 2016. Multi-criteria evaluation of irrigation suitability at national scale: Application of Choquet Integral. *International Journal of Digital Earth*-Submitted

Abstract

In the context of ensuring global and regional food security as well as increasing food production to support the growing world population, irrigation suitability assessment is of paramount importance. This study aims to evaluate the robustness of the Choquet integral (CI) in multi-criteria assessment of surface irrigation suitability for Nigeria. Digital soil mapping products were combined with other landscape data and used for irrigation suitability assessment. Evaluation criteria used include available water capacity (AWC), soil texture, elevation, slope gradient, potential evapotranspiration (PET), aridity index (AI), proximity to perennial water source, proximity to paved roads and proximity to urban markets. Fuzzy membership functions (FMF) were fitted to the criteria values to enable continuous evaluation. The FMF values and weights of the criteria were aggregated using the CI and the weighted sum (WS) aggregation methods to obtain a single irrigation suitability index. Results show that CI is a better aggregator for irrigation suitability assessment than WS. About 3.34×10^6 ha (approximately 4%) of Nigeria is potentially suitable for surface water irrigation. Major limitations are due to topographic and soil attributes. This study indicates a substantial potential to satisfy the significantly increasing demand for food and agricultural products in Nigeria through irrigation agriculture.

Keywords: Irrigation agriculture, Choquet integrals, Fuzzy measures, Food security, Nigeria.

6.1 Introduction

Ensuring global and regional food security as well as increasing food production to support the growing world population rank top among the numerous global challenges in the next quarter century (FAO, 2014). These challenges are overwhelming, especially in SSA, where it is estimated that population level will double by 2050 (FAO, 2009; UN, 2011; Ricker-Gilbert et al., 2014). In the face of enormous food security challenges, agriculture in the SSA will need to undergo a significant transformation in order to match production with increasing food demand amidst the threats posed by climate change (Thornton et al., 2009). To achieve this will require well-coordinated and targeted policies on agricultural intensification through, among others, large and small scale irrigation development (Altchenko and Villholth, 2014). In this light, irrigation suitability assessment and mapping will be pivotal to optimal location of new irrigation developments.

Irrigation suitability assessment involves complex interactions of biophysical, chemical and climatic processes with socioeconomic factors. These processes and factors are in most cases heterogeneous, interdependent and conflicting in nature. Whereas, the biophysical elements tend to be relatively stable, socio-economic factors are dynamic and dependent on the prevailing social, economic and political conditions of an area (Triantafilis et al., 2001; Keshavarzi et al., 2010). Decision-making using such multiple or conflicting criteria can be very subjective and success of such decision-making is to a larger extent dependent on the judgement and expertise of the decision maker (Doumps and Zopounidis, 2011). In many instances these judgements will involve evaluation of an array of alternatives after a careful ranking, sorting or description of decision problems (Roy, 2010). In such situations, where conflicting and disproportionate criteria or choices have to be taken into consideration concurrently, multi-criteria decision evaluation (MCDE) techniques provide a powerful and very robust tool in the hands of decision makers (Mustapha et al., 2011).

Several MCDE techniques have been previously employed in suitability analysis. These techniques are based on simple additive scoring (SAS), weighted average or sum, multi-attribute value technique (MAVT), multi-attribute utility technique (MAUT), ordered weighted average (OWA), fuzzy set theory, and analytic hierarchy process (AHP) (Joerin et al., 2001; Malczewski, 2006; Chen et al., 2010; Zhang and Achari, 2010). Among these approaches, the weighted averaging techniques are the most widely used in suitability analysis (Kordi & Brandt, 2012; Lust and Rolland, 2013) owing to their simplicity and ease of use. However, practical application of these approaches in making decisions involving complex interdisciplinary factors as required for irrigation suitability assessment is limited by their assumption of independency of judgement criteria (Kordi & Brandt, 2012; Mosadeghi et al., 2013; 2015). Such assumption disregards the uncertainty that is usually inherent in human judgements. In addition, they did not allow for interaction among decision criteria and alternatives involved with the aggregation process. In this context, fuzzy set theory techniques have been reported as better alternative methods to deal with these shortcomings, especially in land suitability assessment (Borrough, 1992; McBratney and Odeh, 1997).

Several studies have shown that fuzzy set theory is powerful and very flexible in dealing with the complexities and uncertainties embedded in land suitability assessment (Braimoh et al., 2004; Sicat et al., 2005; Joss et al., 2008; Odeh and Crawford, 2009; Chakan et al., 2012). The performance of fuzzy set techniques in decision making depends largely on the types of membership functions (Zimmermann, 1992). Fuzzy decision-making involves several fuzzy aggregation operators to obtain different types of decision functions. When the array of evaluation criteria and the corresponding weight matrix required for particular decisions are determined, information about the criteria is fused by an aggregation function to determine the overall suitability rating (Soasa and Kaymuc, 2002). Following this approach, conventional fuzzy operators such as t-norms، t-conorms and averaging operators have been employed in soil suitability studies to combine decision criteria into overall suitability index (Chakan et al., 2012). While these aggregation operators can account for the uncertainty involved in human judgement, they do not capture well enough the degree of compensation common to human aggregation ability in the presence of conflicting criteria.

These conflicting criteria present a case where the concept of fuzzy integrals may be appropriate. Among the family of aggregation operators, fuzzy integrals are known to be one of the most robust aggregation functions that allow the fusion of information from several conflicting criteria (Torra and Narukawa, 2006). Fuzzy integral is based on the concept of fuzzy measure, which is a generalization of specific types of averaging aggregation operators (Grabisch et al., 2008). There are several fuzzy integrals: Choquet integral (Choquet, 1954), Sugeno integral (Sugeno, 1974), t-conorm integral (Murofushi and Sugeno, 1991), twofold integral (Torra, 2003), etc. Among these integrals, Choquet integral (CI) is one of the most commonly used for suitability analysis (Wang et al., 2006; Grabisch et al., 2008). It is non-linear, flexible based on either non-additive (Rowley et al., 2015) and/or additive measure. One important feature of CI is the capacity to recognize the vagueness of the decision environment and to account for the interactions among conflicting and correlated criteria (Yang, 2005). CI also considers the degree of satisfaction and/or dissatisfaction of alternatives for each criterion with the help of intuitionistic fuzzy values. In addition, CI allows the quantification of the uncertainty in the aggregation of criteria through sensitivity analysis. CI has been used extensively for MCDE in the field of Engineering, Information Science, Marketing, among others. However, to the best of our knowledge there is dearth of research on the application of CI in irrigation suitability assessment. This study therefore aimed to (i) explore the use of the combination of soil information and landscape attributes for surface irrigation suitability assessment, (ii) evaluate the robustness of Choquet fuzzy integral in a multi-criteria assessment of irrigation suitability assessment at a national scale and (iii) assess the suitability of past decisions on irrigation projects in Nigeria.

6.2 Materials and Methods

6.2.1 Case study

This study is focussed on Nigeria, which is the most populous country in SSA**.** In Nigeria, despite the high demand for food and raw materials, only about 50 percent of the estimated 71.2 million hectares of arable land is currently being utilized (Ayoola, 2009). Although a lot of land is available for agricultural production, the availability of water constitutes a major constraint, especially in the Guinea, Sudan and Sahel Savannah agro-ecological zones of the country (Takeshima and Adesugba, 2014). The over dependency on rain-fed agriculture in Nigeria especially in the semi-humid and semiarid regions and the characteristic erratic rainfall necessitates the practice of irrigation agricultural production where appropriate. In this context, irrigation suitability assessment and mapping would play an important role in ensuring optimal and sustainable agricultural productivity in Nigeria.

6.2.2 Background on fuzzy measures and Choquet integral

6.2.2.1 Fuzzy measures

Before proceeding to the practical aspect of CI, let us look at its theory. The theory of CI found its genesis in the concept of fuzzy integrals which are themselves based on fuzzy measures or capacities interpreted as the generalization of specific weighting vectors as used in the computation of weighted sums (Grabisch et al., 2008). In defining fuzzy measure, let $X = \{x_1, x_2, \dots, x_n\}$ be a finite set of decision alternatives and let $P(X)$ denote the power set of X, or a set of all subsets of X. A fuzzy measure g defined on X is a function: $g: P(X) \to [0,1]$ such that:

$$g(\emptyset) = 0, g(X) = 1. \quad (6.1)$$

$$\text{If } A, B \subseteq P(X) \text{ and } A \subseteq B \text{, then } g(A) \leq g(B). \quad (6.2)$$

If eq (2) is not satisfied, g is called a non-monotonic fuzzy measure. Sugeno introduced the so called λ fuzzy measure satisfying the following additional property: for all $A, B \subset X$ with $A \cap B = \emptyset$,

$g(A \cup B) = g(A) + g(B) + \lambda g(A)g(B)$, for some fixed $\lambda > -1$. (6.3)

The value of λ can be found from$g(X) = 1$, which is equivalent to solving equation:

$$\lambda + 1 = \prod_{i=1}^{n}(1 + \lambda g_i). \tag{6.4}$$

Let $A = \{x_i, x_{i+1}, \ldots, x_n\}$. If q is a λ fuzzy measure, the value of $g(A_i)$ can be computed recursively as:

$$g(A_n) = g(\{x_n\}) = g_n, \tag{6.5}$$

$$g(A_i) = g_i + g(A_{i+1}) + \lambda g_i g(A_{i+1}) \qquad \text{for } 1 \leq i < n. \tag{6.6}$$

In multi-criteria decision making, $g(A_i)$ can be viewed as the importance of the criterion or expert set A. Thus, in addition to the usual weights on the experts and on the criteria taken separately, weights on any combination of the criteria and the experts are also defined.

6.2.2.2 Choquet fuzzy integral

The CI is a fuzzy integral based on λ fuzzy measure that provides alternative computational scheme for aggregation information. To define CI we assume $h(x_1), h(x_2), \ldots, h(x_n)$ are a collection of input resources of h , and if g is a λ fuzzy measure, then we can construct a CI as:

$$\int_x h(x) \circ g(x) \qquad [6.7]$$

Alternatively, assuming X is a finite and discrete set, the Choquet fuzzy integral can be computed as follows:

$$E_g(h) = \sum_{i=1}^{n}[h(x_i) - h(x_{i-1})]\, g(A_i) \qquad [6.8]$$

where $h(x_1) \leq h(x_2) \leq \cdots \leq h(x_n)$, and $h(x_0) = 0$. Another computation formula for the finite set case can also be represented by

$$E_g(h) = \sum_{i=1}^{n} h(x_i)[g(A_i) - g(A_{i+1})] \quad [6.9]$$

6.2.2.2.1 Importance index of evaluation criteria.

One important attribute of CI is the ability to gauge the importance of each evaluation criterion as well as the combination of criteria for the decision process. This is achieved using the concept of Shapley index (Shapley, 1953). Again, let g be a fuzzy measure and $X = \{x_1, x_2, ..., x_n\}$ be the set of evaluating criteria. Then the Shapley index ϕ_i for every input $x_i \in X$ can be defined as:

$$\phi_g(i) = \sum_{A \subseteq X/[x_i]} \frac{(n-|A|-1)!|A|!}{n!}[g(A \cup \{x_i\}) - g(A)] \quad [6.10]$$

The Shapley index ϕ_i can be interpreted somewhat in two ways: (i) as the average value of the contribution of criteria x_i in all coalitions of criteria and (ii) as the true representation of the sharing of the total amount $g(X)$ since it must satisfy the condition $\sum_{i=1}^{n} \phi_g(i) = 1$ i.e., the sum of importance degrees of all coalitions is a constant.

6.2.2.2.2 Interaction index of evaluating criteria.

Another important capability of CI that differentiates it from the other aggregators is its ability to cater for interactions among the evaluating criteria through the interaction indices. For instance, given a set of evaluating criteria $X = \{x_1, x_2, ..., x_n\}$, the interaction index for every individual set $x_i \in X$ can be defined as:

$$I_g(A) = \sum_{B \subseteq X/A} \frac{(n-|B|-|A|)!|B|!}{(n-|A|+1)!} \sum_{C \subseteq A} (-1)^{|A/C|} g(B \cup C) \qquad [6.11]$$

This measure is true for all coalitions of evaluation criteria whereby $I_g(A) \in [-1,1]$. For easier interpretation, the interaction index $I_g(ij)$ for each pair of criteria $A = \{x_i, x_j\}$ is commonly used. A pair of criteria x_i and x_j is said to have a positive interaction (complement) if $I_g(ij) > 0$. This implies that the satisfaction of both criteria is necessary for an overall satisfaction of the decision process. On the contrary, a negative interaction (correlation) exists between x_i and x_j if $I_g(ij) < 0$. This implies that both criteria are substitutive and it is therefore sufficient to satisfy either of them to get overall satisfaction. If $I_g(ij) = 0$, then x_i and x_j are said to have little or no interaction (independence) existing between them.

6.2.3 Selection and processing of evaluation criteria

Several parameters are required for surface water irrigation suitability assessment, particularly those related to the properties that govern irrigation water availability, erosion or sedimentation, drainage, salinity, market outlets and accessibility, length of growing season, among others. For the purpose of this study, five main decision criteria, namely: topography, climate, soil properties, socioeconomic and hydrology, were selected. Based on these decision criteria, nine evaluation criteria were selected for potential irrigation suitability assessment. The selection of the evaluation criteria was largely limited by data availability as well as maximum evaluation input for the CI software used for this study (Takahagi, 2005b).

6.2.3.1 Climatic variables

Climatic variables are among the major factors that determine the potential irrigation suitability of an area. Because of their systemic variations across a given landscape, climatic variables will decide the regional irrigation needs. For instance, there will be greater need for irrigation to complement crop water needs in drier arid or semi-arid regions than in humid regions. Climatic variables used in this study include Aridity Index (AI) and annual Potential Evapo-Transpiration (PET). The Aridity index is mostly expressed as a generalized function of precipitation, temperature, and/or PET (UNEP, 1997) while PET measures the ability of the atmosphere to remove water from the earth surface through Evapo-Transpiration (ET) processes. Both attributes can give an indication of the level of water deficit as well as the dryness of the climate of a particular area (Kumbhar et al., 2014). AI values for instance increase in humid conditions, and decrease in arid conditions. Global AI and global PET geo-spatial datasets were obtained from the Consortium for Spatial Information (CGIAR-CSI) GeoPortal (Trabucco and Zomer, 2009) and clipped for Nigeria. The global AI and PET were modelled at 1km resolution using the "WorldClim" global climate data (Hijmans et al. 2005) as input parameters.

6.2.3.2 Topographic indices

Suitability of a land for surface water irrigation depends largely on its topography. Topographic features generally affect irrigation efficiency, drainage pattern, erosion intensity and cost of land development (Ali, 2010). Topography is also a good proxy for representing subsurface water flow paths as well as soil–water storage dynamics of a given area (Lanni et al., 2011). Therefore, in this study two major topographic indices (elevation and slope) were included in the evaluation criteria. Elevation data was obtained for our study area using the SRTM 3 arc (90 m) digital elevation model (DEM) while slope gradient was derived from the DEM using the Spatial Analyst Toolbox in ArcGIS 10.2 (Reuter and Nelson, 2009).

6.2.3.3 Soil attributes

Soil attributes are important factors to be considered in decisions on irrigation development (Frenken and Faurès, 1997) in that they affect the crop growth as well as drainage of a particular area. Soil available water capacity (AWC) and clay content were the soil parameters considered for this study. Clay content was obtained from previous digital soil mapping study of the area (Akpa et al., 2014) while AWC was estimated for Nigeria using the pedotransfer (PTF) of Minasny and Hartemink (2011) thus:

$$\theta_{-33kpa}(\%) = (56.5 - 7.49BD - 0.34Sand) \quad [6.12]$$

$$\theta_{-1500kpa}(\%) = 7.95 + 0.86 * OC + 0.4 * Clay - 0.004 * (Clay - 37.7)^2 \quad [6.13]$$

$$AWC = \theta_{-33kpa} - \theta_{-1500kpa} \quad [6.14]$$

where, AWC is soil available water capacity, BD is soil bulk density in mass per unit volume, Clay is clay content in percent mass, Sand is sand (particles 50–2000 μm) content in percent mass, OC is organic carbon content in percent mass while θ_{-33kpa} and $\theta_{-1500kpa}$ are soil moisture content at field capacity and permanent wilting point respectively. The input variables used in the above PTFs were obtained from previous chapters in this thesis: BD and OC (chapter 4) while clay and sand contents were from chapter 3.

6.2.3.4 Proximity to river and road networks

The shapefiles of stream network and road network of the study area were obtained from Diva_GIS website (http://www.diva-gis.org/gdata). These shapefiles were overlaid by the DEM map and the proximity of each pixel to the nearest river and road was determined separately using the Analysis Toolbox in ArcGIS 10.2.

6.2.3.5 Proximity to urban market

Proximity of market outlets for agricultural crops is an important factor for agricultural planning and irrigation development. Generally, crops of high market value and high transportation costs are grown nearest to the market while less perishable crops with lower production and transportation costs are grown farther away. For the purpose of this study, potential market outlets were determined using the location information of major towns with a population of at least 50,000 according to the 2006 population census data of Nigeria following Worqlul et al. (2015). Using the DEM map, the proximity of each location was determined the Analysis Toolbox in ArcGIS 10.2.

6.2.4 Derivation of fuzzy membership functions on the primary input variables

Decision criteria used in MCDE are usually from different domain and with different scale of measurement. Therefore one very crucial step in MCDE is the standardization of evaluation criteria. Scales of 0 to 1, 0 to 5, 0 to 10, 0 to 100, among others are usually used for standardization. However, standardization is achieved in CI aggregation technique via the fitting of fuzzy membership functions (FMF) to input variables. FMF is an extension of the classic binary logic, with the capability of defining sets without sharp boundaries and allowing for partial degree of membership (Borrough, 1993). FMF transforms the input data to the real unit interval of 0 to 1 based on the possibility of being a member of a specified set. The value 0 means non-membership of the fuzzy set; the value 1 means full membership of the fuzzy set. Any value between 0 and 1 characterize a partial membership of the set.

During the “fuzzification” of the evaluation criteria, the choice of the FMF was guided by expert knowledge and based on the strength of each input in the decision process. McBratney and Odeh (1997) discussed a range of possible membership function applicable to soil suitability assessment. For the purpose of this study, FMFs were established for each evaluation criterion discussed in Section 6.2.3 above using the fuzzy membership tool of the Spatial Toolbox in ArcGIS 10.2. The fuzzy “MS large” and fuzzy “MS small” functions which implement the sigmoid FMF algorithm were used in the fuzzification of input rasters. Fuzzy “MS large” function defines a fuzzy membership based on the mean and standard deviation, with the larger values having a membership closer to 1. “MS small” function, uses similar approach but with smaller values having a membership close to 1 as shown in Fig. 6.1.

6.2.5 Weighting of evaluation criteria

The determination of weights of evaluation criteria is a crucial step in every multi-criteria analysis ((Feizizadeh and Blaschke, 2013). Therefore, in this study weights of the selected evaluation criteria were assigned using the pairwise comparisons approach provided in a CI interface developed by Takahagi (2005b). In the light of the pairwise comparison, each criterion was matched head-to-head based on the relative contribution to the overall irrigation suitability index, to obtain a comparison matrix. The head-to-head rating of the decision criteria was done using local expert opinions on their relative importance to surface irrigation development. Based on expert opinions, a scale of importance in the range of 1 to 9 (see Table 6.1) were assigned to each criterion with the value of 9 corresponding to absolute importance and 1 representing equal importance. Since human judgement is subjective, a consistency index (CR), which is a reflection of how consistent the decision maker is in placing importance on the evaluation criteria was obtained for each pairwise comparison matrix.

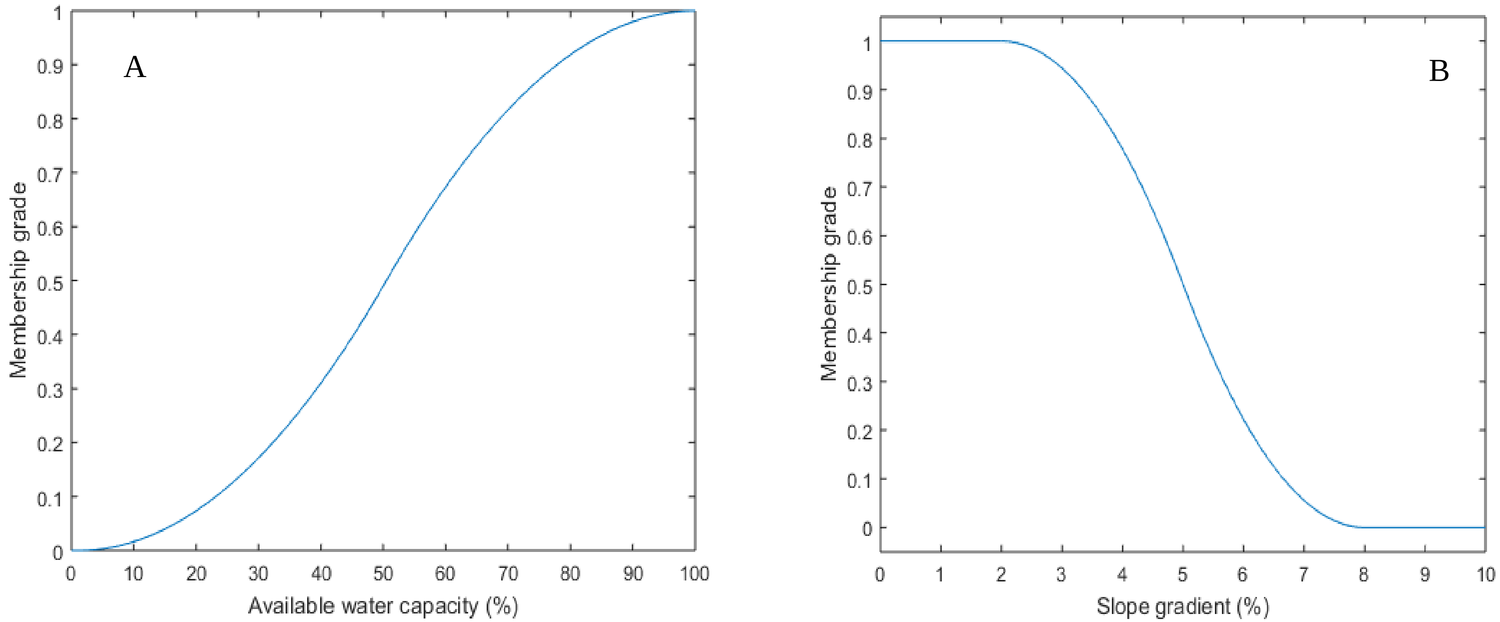


Figure 6.1 Examples of fitted membership grades of evaluation criteria using sigmoid membership functions.

Table 6.1 Fundamental scale used in pairwise comparison based on Saaty (2008)

Intensity of Importance	Qualitative Definition	Explanation
1	Equal importance	Two factors contribute equally to the objective
3	Somewhat more important	Experience and judgement slightly favour one item over the other
5	Much more important	Experience and judgement strongly favour one item over the other.
7	Very much more important	Experience and judgement very strongly favour one item over the other. Its importance is demonstrated in practice.
9	Absolutely more important	The evidence favouring one item over the other is of the highest possible validity.
2,4,6,8	Intermediate values	When a compromise is needed
Reciprocals of above	If item i has one of the above non-zero numbers assigned to it when compared with item j, then j has the reciprocal value when compared with i	A reasonable assumption

6.2.6 Irrigation suitability assessment

The surface irrigation potential of our study area was determined by weighting the evaluation criteria as discussed previously and aggregating each using two aggregation methods: CI and weighted sum (WS). As shown in Fig.6.2 and following Eqs.[6.1-6.11], the required inputs for the use of CI to aggregate suitability indicators into a single suitability index are the degrees of memberships of the evaluating criteria or suitability indicators (h) and the relative importance of their weights (g), the choice of interaction index and the identification standard methods. We envisaged that the choice of parameters required in fitting CI based on the steps illustrated in Fig 6.2 will have varying effect on the outputs. Therefore a sensitivity analysis was done to evaluate the robustness of the suitability index to changes in the parameter settings.

Prior to fitting of the CI, 1000 sample locations were randomly selected using conditioned Latin hypercube sampling (cLHS) scheme (Minasny and McBratney 2006) based on five covariates; DEM, slope, aspect, mean annual rainfall, and land use. Fuzzy membership values (FMVs) of the criteria were then extracted at these locations and aggregated using CI to obtain a single suitability index at each point. The CI was implemented using the CI interface developed by Takahagi (2005b). The outputs of the analysis, which are the suitability indices at the 1000 locations, were then interpolated across Nigeria using Random Forest Model in R environment (R Development Core Team, 2014). For the spatial interpolation we used covariates such as topographic wetness index, stream power index, flow accumulation, elevation, slope and topsoil sand content that exhibit significant correlations with suitability indices obtained at the 100 locations. The weighted sum technique was carried out in ArcGIS environment by aggregating the map layers of the “fuzzified” evaluation criteria using the map overlay tool.

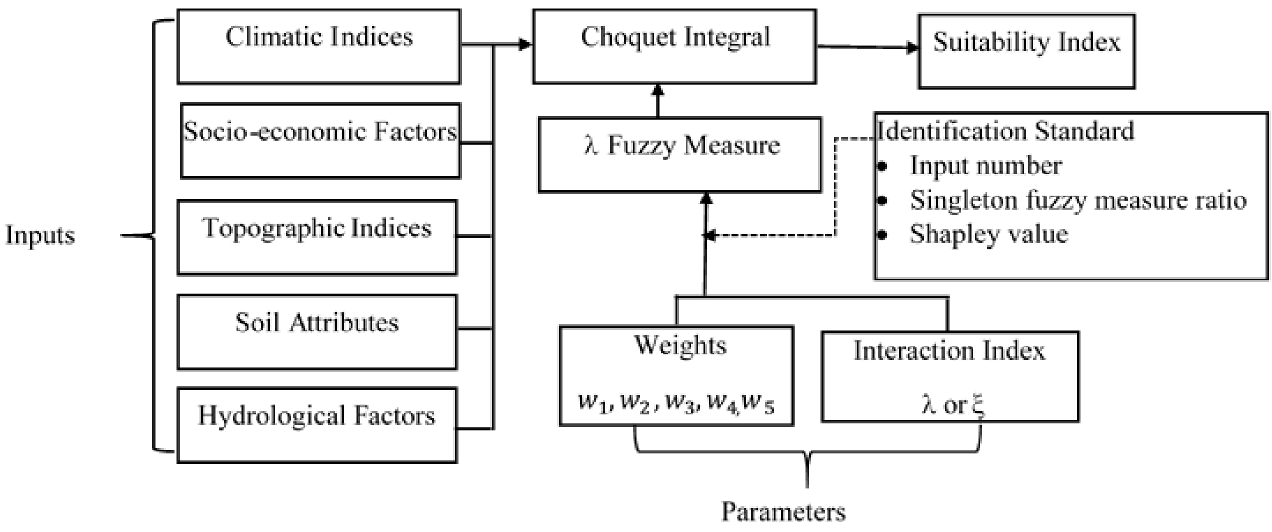


Figure 6.2 Schematic illustration of the application of Choquet integral with λ fuzzy measure in suitability assessment. Adopted and modified from Takahagi 2005a

6.2.7 Estimation of most suitable irrigation areas across different regions

Each of the suitability index maps computed using the two aggregating approaches of CI and WS described above, was then multiplied by a restriction map. Restriction areas are those areas considered permanently non-suitable for irrigation and usually comprise of water bodies, built-up areas, forestlands and existing road networks which (Worqqlul et al., 2015). The final irrigation suitability maps were then classified into four irrigation suitability classes following FAO (1976). A user defined threshold value of 0.69 was employed in choosing highly suitable areas. The raster maps of highly suitable areas for surface irrigation were then converted to polygons in ArcGIS environment. Number of continuous areas (based on the polygons) suitable for small, medium and large irrigation across the various agro-ecological and geopolitical zones in the study area were then determined following (Worqlul et al., 2015). We included geopolitical zones in this analysis because most of the national agricultural developmental policies in Nigeria are currently carried out along the various geopolitical zones.

6.3 Results

6.3.1 Spatial distribution and interaction among evaluation criteria

The spatial distribution of the evaluation criteria used for the multi-criteria assessment of our study area is presented in Fig. 6.3. There is a systemic variation of Aridity index and PET from the coast towards the inland of Nigeria. Aridity index decreased gradually from the southern region northward, with a slight increase around the Jos Plateau in the central coast. In contrast, PET increased gradually from the south coast to the northern part of the country. The northern part of Nigeria is dryer than the southern region as indicated by their aridity index and PET. In terms of topographic indices, the landscape of the study area is relatively low lying with flat slope in the southern coastal region and around the fringes of the northern region. However, highlands with steep slope gradients interspersed the western, eastern borders and middle belt regions. Generally, soils of the study area are characterised by relatively coarse texture with a concomitant low to medium water holding capacity.

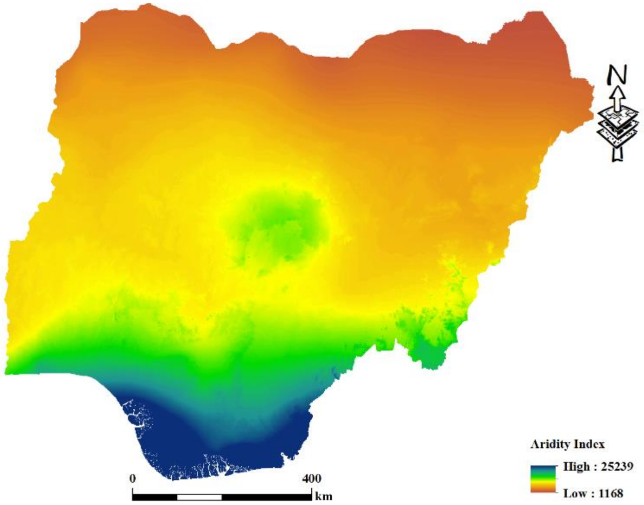
The Pearson product moment correlation coefficient matrix of both the raw values and fuzzy grades of the evaluation criteria are shown in Tables [6.2 to 6.3]. The correlation coefficient matrix indicates a moderate to strong correlation among the evaluation criteria. For instance, Aridity index is negatively correlated with almost all the other factors. In contrast, PET is positively correlated with most factors with the exception of Aridity index and clay content. As expected, there is a strong negative correlation between aridity index and PET. Also, elevation shows a strong positive correlation with slope, proximity to road and proximity to urban markets.

6.3.2 Weighting of evaluation criteria for irrigation suitability

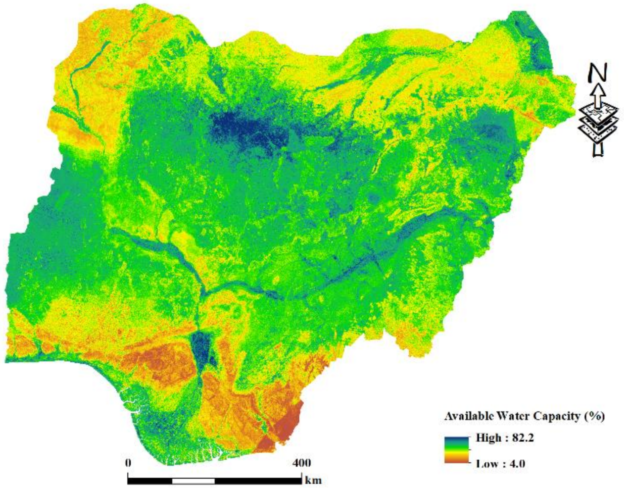
The result of the pairwise comparisons matrix for the nine evaluation criteria based on Table 6.1 is presented in Table 6.4. It is evident that proximity to the river with a weight of 26% is the most important criteria for surface irrigation suitability assessment in the area. This is followed closely by elevation with 21 % weight and slope gradient with 14.4 % weight. In addition, soil attributes and climatic variables were considered of moderate important proximity to market outlet with 3% weight is the least important. As mentioned in the previous section, the head-to-head weighting of the decision criteria was done using literature search and local expert opinions on their relative importance to decisions on surface irrigation development. In this study, the CR for the pairwise comparisons matrix used in criteria weighting is 0.06.

6.3.3 Effect of parameter settings on Choquet integral suitability index

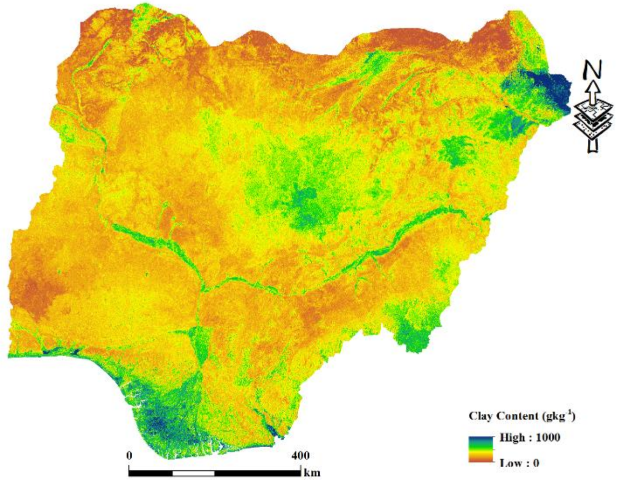
We evaluated the effect of different interaction indices (ξ and λ values) on the output of CI aggregation (results not shown here for lack of space). At a low interaction index **(**ξ =0.2, λ=15), that is when the decision maker places importance strongly on the balance among evaluation criteria (Takahagi, 2005b**),** there is very large distribution of low to marginal irrigation suitability across the study area. At high interaction index (ξ =0.8, λ=-9375), that is when the decision maker do not place importance strongly on the balance among evaluation criteria (Takahagi, 2005b**),** the distribution of suitability index narrowly ranged and predominantly on the high side of irrigation suitability. However, when there is moderate interaction index (ξ =0.49, λ=0.05), that is when the decision maker places weak importance on the balance among evaluation criteria; there is a wider distribution of the suitability index. Therefore the optimum interaction degree is about when the interaction index, ξ equals 0.49 and λ equals 0.05.



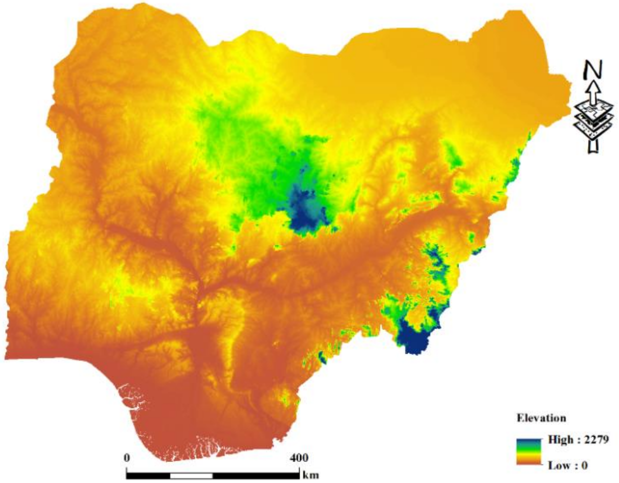
(a) Aridity Index



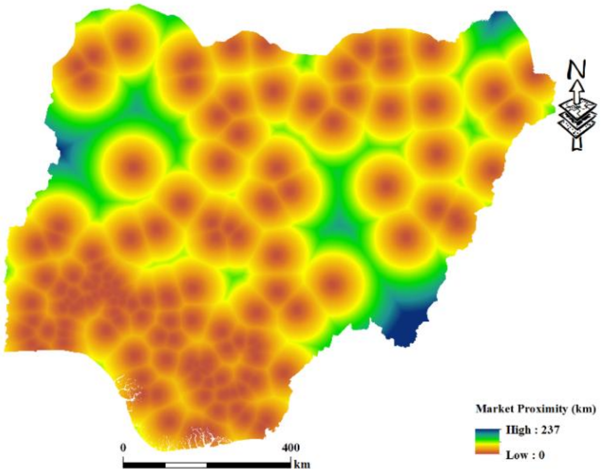
(b) Available water capacity



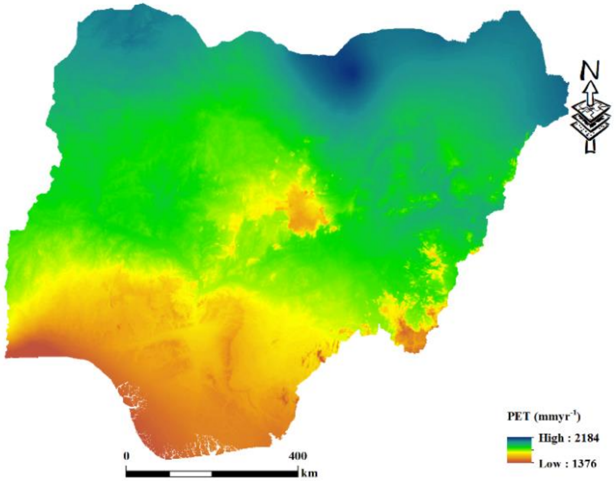
(c) Clay content



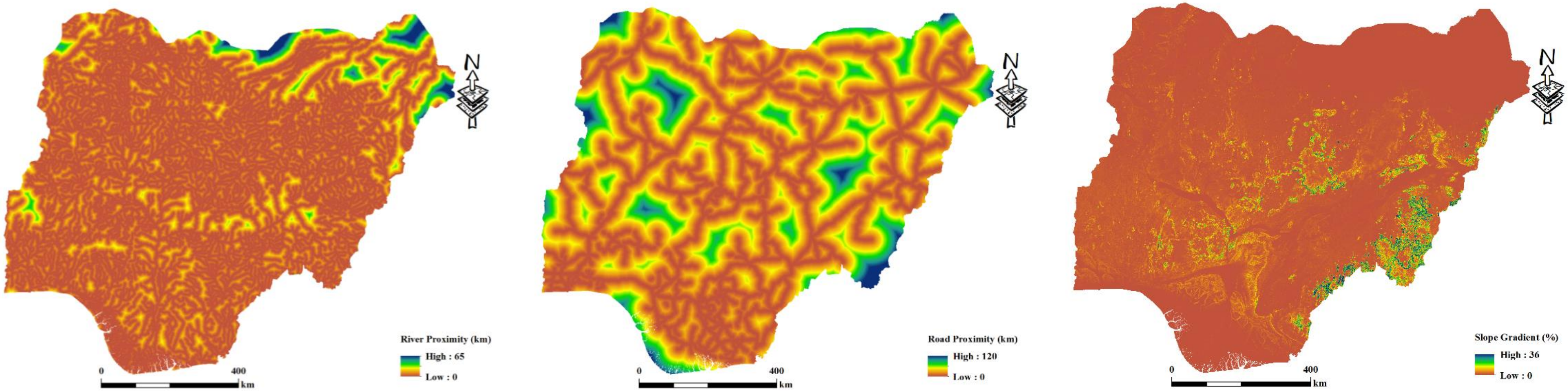
(d) Elevation



(e) Market proximity



(f) Potential Evapo-Transpiration



(g) River proximity (h) Road proximity (i) Slope gradient

Figure 6.3 Evaluation criteria (a-i) used for irrigation suitability assessment.

Table 6.2 Pearson correlation coefficient matrix of evaluation criteria used in irrigation suitability assessment.

	AWC	Clay	Slope	Aridity_I	PET	elevation	Road_P	River_P	Market_P
AWC	1.000								
Clay	0.171***	1.000							
Slope	-0.180***	0.034NS	1.000						
Aridity_I	-0.156***	0.290***	0.127***	1.000					
PET	0.124***	-0.205***	-0.190***	-0.906***	1.000				
elevation	0.192***	0.124***	0.347***	-0.287***	0.191***	1.000			
Road_P	0.067*	0.047NS	0.135***	-0.077**	0.089**	0.245***	1.000		
River_P	-0.123***	-0.026NS	-0.038NS	-0.235***	0.266***	0.005NS	0.056NS	1.000	
Market_P	0.130***	0.052NS	0.196***	-0.165***	0.195***	0.333***	0.571***	0.032NS	1.000

*, p value<0.05; **, p value<0.01; ***, p value<0.0001; NS, Not significant; AWC, available water capacity; Aridity_I, aridity index; PET, potential evapo-transpiration; Road_P, road proximity; River_P, river proximity; Market_P, market proximity.

Table 6.3 Pearson correlation coefficient matrix of fuzzy grades of evaluation criteria used in irrigation suitability assessment.

	AWC	*Clay*	*slope*	*Aridity*	*PET*	*Elevation*	*Road_P*	*River_P*	*Market_P*
AWC	1.000								
Clay	0.188***	1.000							
slope	0.238***	-0.066*	1.000						
Aridity	-0.181***	0.293***	-0.091*	1.000					
PET	-0.169***	-0.055NS	0.250***	-0.514***	1.000				
Elevation	-0.237***	-0.228***	0.370***	0.123***	0.114***	1.000			
Road_P	-0.009NS	-0.058NS	0.110***	0.066*	-0.065*	0.121***	1.000		
River_P	-0.133***	-0.043NS	0.143***	0.054NS	0.115***	0.355***	0.169***	1.000	
Market_P	-0.051NS	-0.060NS	0.176***	0.140***	-0.065*	0.195***	0.475***	0.162***	1

*, p value<0.05; **, p value<0.01; ***, p value<0.0001; NS, Not significant; AWC, available water capacity; Aridity_I, aridity index; PET, potential evapo-transpiration; Road_P, road proximity; River_P, river proximity; Market_P, market proximity.

Table 6.4 Pairwise comparison matrix and weights of evaluation criteria.

	AWC	Clay	Aridity_I	N_River	N_Market	N_Road	PET	Elevation	Slope	Weights
AWC	1	2	4	1/2	5	4	4	1/3	1/2	0.132
Clay	1/2	1	3	1/3	4	3	3	1/3	1/2	0.094
Aridity_I	1/4	1/3	1	1/5	3	2	1	1/4	1/4	0.049
N_River	2	3	5	1	7	6	4	2	3	0.262
N_Market	1/5	1/4	1/3	1/7	1	1	1/2	1/5	1/4	0.028
N_Road	1/4	1/3	1/2	1/6	1	1	1/3	1/4	1/4	0.031
PET	1/4	1/3	1	1/4	2	3	1	1/5	1/4	0.048
Elevation	3	3	4	1/2	5	4	5	1	2	0.212
Slope	2	2	3	1/3	4	4	4	1/2	1	0.144

Consistency index = 0.06; AWC; available water capacity, Aridity_I; aridity index, PET; potential evapo-transpiration, N_Road; road proximity, N_River; river proximity, N_Market*;*

6.3.4 Spatial distribution patterns of potential surface irrigation suitability.

The spatial distribution of the potential surface irrigation suitability areas based on the outcome of CI and WS is presented in Fig. 6.4. Comparing the effect of equal and varying weights of evaluation criteria on suitability index, it could be seen that assigning equal weights to all the evaluation criteria results in low to marginal suitability index across the study area regardless of whether CI or WS is employed (Figs. 6.4a and 6.4c). However, the allocation of varying importance to the different evaluation criteria results in a moderate to high suitability index (Figs. 6.4b and 6.4d). The suitability index estimated by WS ranged from 0.22 to 0.88 (mean = 0.64, SD = 0.065) while those of CI ranged between 0.35 and 0.82 (mean = 0.65, SD = 0.054). CI results in smaller areas with low suitability index (less than 0.45) compared to WS. On a closer look at Fig. 6.4d, one could see that there is an undue interference of extreme values of evaluation criteria, especially river proximity on the outcome of WS aggregation technique. In contrast, CI looks more robust as it averages out the impact of these extreme values. Additionally, CI better captured the higher irrigation suitability in the vicinity of major rivers (see Fig. 6.4c) than captured by WS (see Fig. 6.4d). Intuitively, CI performed better than WS in the study area. Therefore, we shall limit our reports on other results and discussions to only the output of CI analysis.

6.3.5 Regional variation of potential surface irrigation suitability

The potential surface irrigation suitability map based on the best aggregator (CI) was multiplied by the restriction map of the study area as discussed earlier and then further optimized using a user defined threshold value of 0.68 to capture best suitable areas for irrigation. The optimized suitability maps were used to estimate parcels of suitable irrigation lands across the various regions in the study area (see Table 6.5). Overall, about 3.34 x 10^6 ha of land is physically suitable for surface irrigation development in Nigeria. The northern region (comprising of northcentral, northeast and northwest zones) collectively gave a significantly higher proportion of suitable areas (about 76% of the entire area) than the southern region (southeast, southsouth and southwest) with only about 24% suitability. Looking at irrigation suitability of the individual zones, the northcentral zone has the largest proportion of suitability areas, constituting about 29 % suitability of the entire country. This is followed closely by the northeast zone (25.5%) and the northwest zone (21.9 %). The relative proportion of suitable areas in the south east, southsouth and southwest regions, with respect to the suitability of entire country, is about 6.2, 12.4 and 5.2 percent respectively.

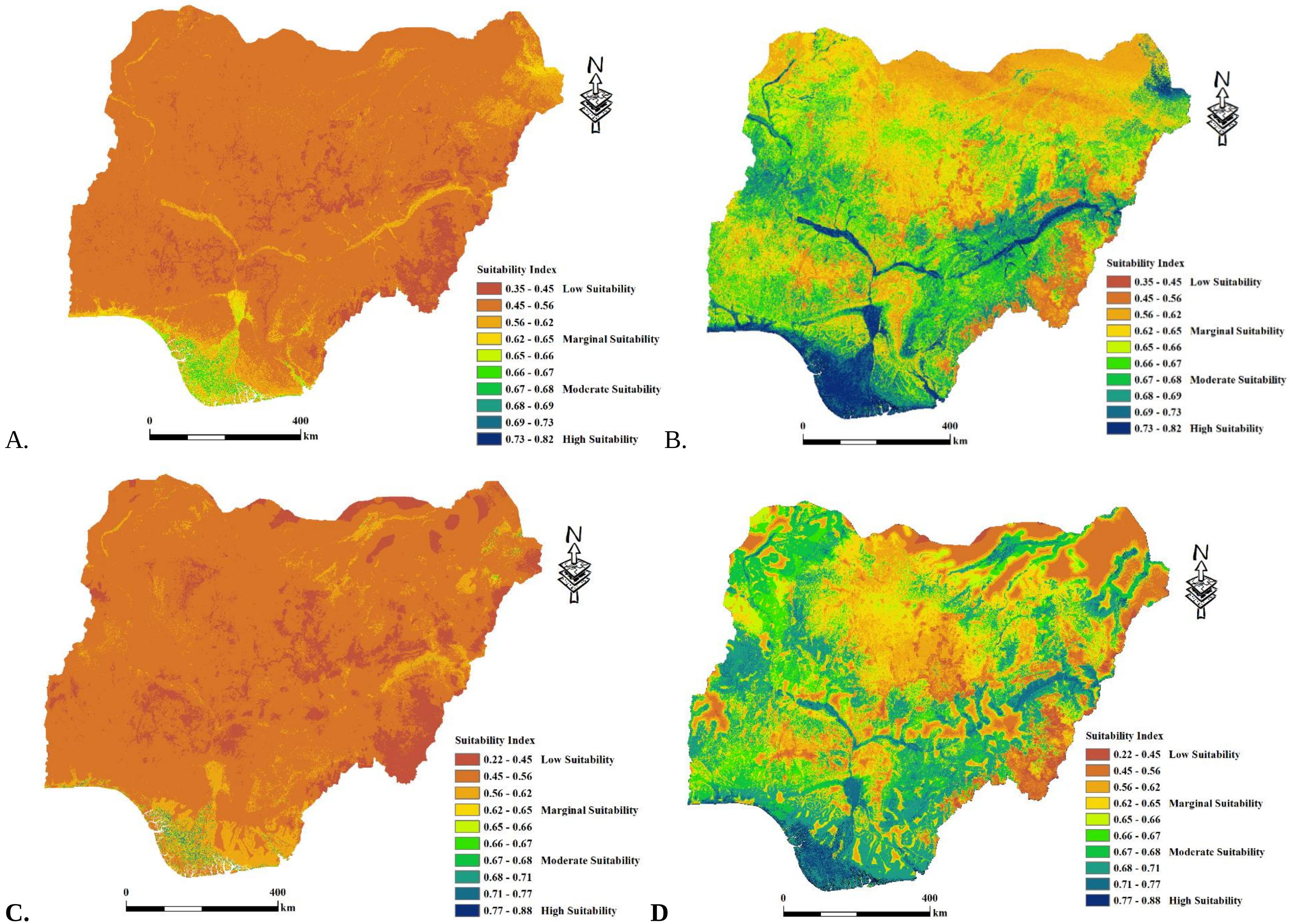


Figure 6.4 Irrigation suitability using equal and varying weights based on Choquet integral (A-B) and weighted sum (C-D).

Table 6.5 Potentially suitable areas for surface irrigation across the various geopolitical zones of Nigeria based on Choquet Integral.

Regions	Most Suitable area (ha)	Number of areas suitable for large scale surface irrigation (>3000 ha)	Number of areas suitable for medium scale surface irrigation (>200 and <3000 ha)	Number of areas suitable for small scale surface irrigation (<200 ha)	Percentage of potentially suitable area (%)
			Geopolitical zones		
Northcentral	963500	37	366	2800	28.8
Northeast	851700	33	314	1796	25.5
Northwest	733100	28	435	2105	21.9
Southeast	208540	7	106	534	6.2
Southsouth	414700	21	200	816	12.4
Southwest	169800	5	109	523	5.2
Total	3341600	131	1530	8574	100
Proportion of total (%)		1.3	14.9	83.8	
			Agro-ecological zones		
SHS	358700	11	102	590	10.7
SDS	121500	2	93	483	3.7
NGS	73000	2	32	451	2.2
SGS	438400	22	187	1233	13.1
DS	1233000	44	518	3135	36.9
HF	1117000	48	622	2787	33.4
Total	3341600	129	1554	8679	100
Proportion of total (%)		1.2	15	83.8	

DS, Derived Savannah; HF, Humid Forest; NGS, Northern Guinea Savannah; SDS, Sudan Savannah; SGS, Southern Guinea Savannah; SHS, Sahel Savannah

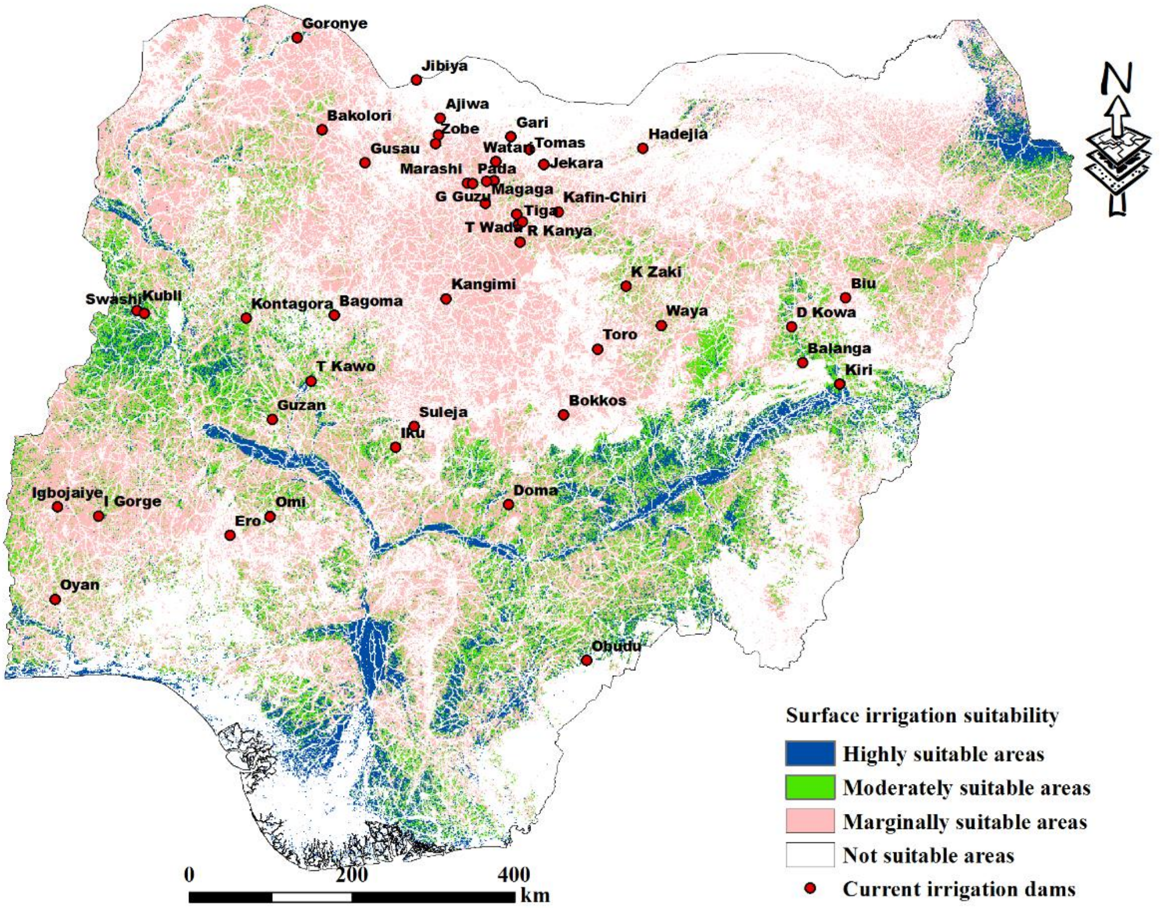


Figure 6.5 Suitability of existing public irrigation sites in Nigeria.

6.3.6 Evaluating past decisions on irrigation projects in Nigeria.

To evaluate past decisions regarding the siting of irrigation projects we obtained location coordintates of the existing irrigation dams in Nigeria from the AQUASTAT database (Frenken, 2005). These were overlaid on the optimized irrigation suitability map as described in previous session to assess whether previous irrigation projects were properly situated (see Fig.6.5). Results indicate that only a very few of the existing public irrigation dams were sited in highly suitable surface irrigation areas. Notable of these are irrigation projects within the Sokoto rima basin of the northwestern part of the country: Goronye dam, Swashi and Kubil Dam. Others are

Kiri and Doma dams in northeast and northcentral regions of Nigeria. Notwithstanding, a good number of past irrigation dams were located in the moderately suitable irrigation areas. In addition, almost 50% of irrigation dams under the Hadeija Jama'are river basin development authority are in the marginally suitable areas. Furthermore, a few irrigation dams like the Jibiya dam in the border area of Nigeria and Niger, Bokkos dam on the central plateau and Ero dam in mid-western part of Nuigeria are somewhat located in currently not suitable areas.

6.3.7 Sensitivity analysis of irrigation suitability index

The sensitivity analysis of computed suitability index at selected sample locations based on different interaction indices is shown in Fig. 6.5. On a closer look, one could see that location 200 is the most suitable for surface irrigation when ξ (xi values) > 0.35, that is, if the decision maker do not place importance strongly on the balance among evaluation criteria (Takahagi, 2005b**)**. In contrast, when ξ (xi values) < 0.2, that is, if the decision maker places importance strongly on the balance among evaluation criteria (Takahagi, 2005b**)**, location 5 is the most suitable for surface water irrigation. Locations 50 and 150 show equal suitability when ξ >0.45 while location 50 is the most suitable among the two locations when ξ <0.45. At all ξ values, location 450 is the least suitable for surface water irrigation among the selected locations. Overall, there is a clearer distinction between the suitability index of the selected locations at ξ <0.5 compared to higher ξ values. This is an indication that the best ξ for CI suitability index evaluation in the study area is most likely to be a little below 0.50.

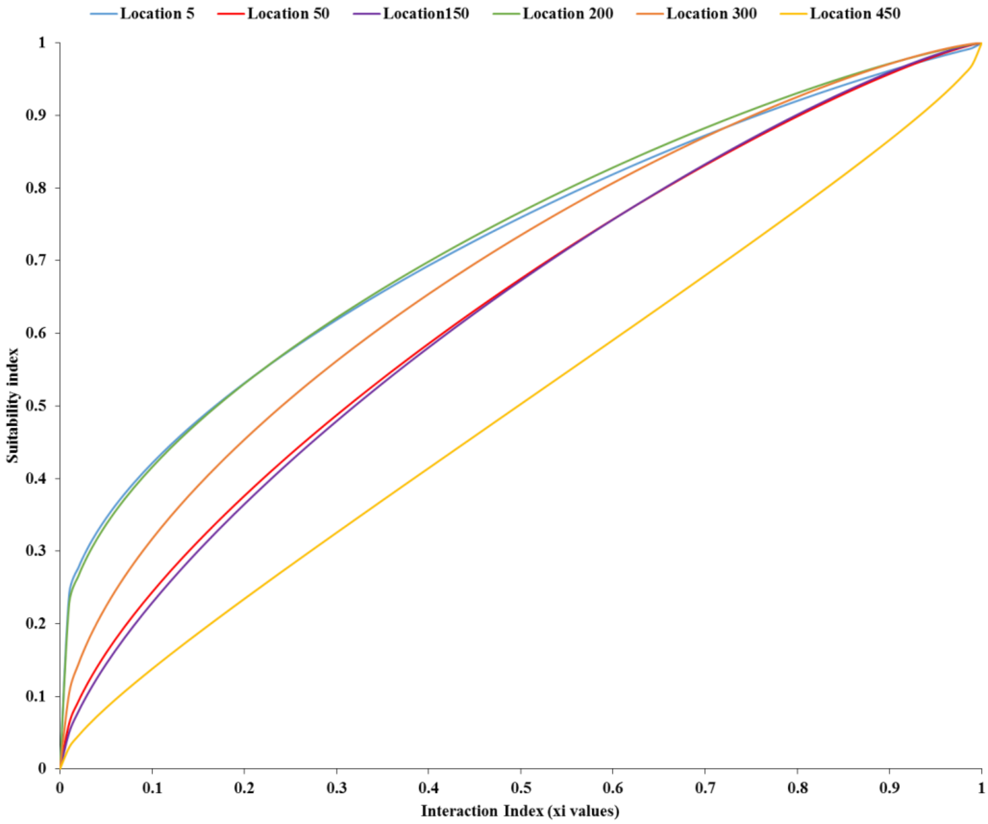


Figure 6.6 Sensitivity analysis of irrigation suitability index at selected sample locations.

6.4 Discussion

6.4.1 Distribution and interrelationship among criteria

There is a systemic variation of both the aridity index and PET across the study area compared to other variables. This implies that these two variables will influence decisions on the regional irrigation needs across the study area while the other factors may influence decisions on local irrigation needs. The northern region of the study area is significantly dryer than the southern region as indicated by lower aridity index and higher PET. This is justified by the shorter length of rainy season as well as high evaporative demand in the northern region compared to south coast. Variability in rainfall regimes and atmospheric evaporative demands have been reported as the most prominent environmental factors responsible for the temporal dominance of drier soils in the semi-arid environment than the sub-humid areas (Lauenroth and Bradford, 2006; 2009). The dryness of the northern region implies a greater need for irrigation to support agricultural development in this area compared to its southern counterpart.

The topography of the study area is relatively uniform to a large extent except for steep slope gradients along the eastern borders, mid-western and around Jos Plateau in the middle belt region. This implies that the development of surface irrigation scheme in our study area will be cost-effective to some extent, since topography significantly influences the initial cost of surface irrigation development (Ali, 2010). Generally, the coarse texture of soils in most part of the northern Nigeria suggests that drainage may not pose much challenge to surface irrigation in that area. However, low nutrient availability may constitute major challenge to irrigation agriculture in the long run. In addition, the denser distribution of urban markets in most part of the southern regions than their northern counterpart implies that while the north may have higher potential for irrigation farming, most of the crop produce will need to be conveyed to the southern region where there are larger market outlets for them.

The significant correlation among the selected evaluation criteria is an indication of possible interaction in the form of synergy or redundancy between these criteria when used for multi-criteria analysis. However, these interactions can only be mild considering that there only exist a weak to moderate relationship among most criteria except for between elevation and slope, elevation and market proximity, PET and AI, road and market proximities. The existence of interaction among the evaluation criteria used in this study (although not too strong in many cases) lends credence to the robustness of CI for criteria aggregation in this study.

6.4.2 Weighting of evaluation criteria for irrigation suitability

The result of our pairwise comparisons matrix for the nine evaluation criteria is relatively consistent, in that our reported CR of 0.06 is below the acceptable threshold of 0.1 (Chen et al., 2013). Also, the trend of the assigned weights is similar to that reported by Worglul et al. (2015) for similar irrigation study in the Lake Tana Basin of Ethiopia. However, our results are invariant with theirs in terms of the second most important criteria as they have reported proximity to road as the second most important criteria for surface water irrigation development. This could be attributed to the smaller coverage of their study area compare to ours. Moreso, in Nigeria, farming activities are usually carried out in the rural areas with little or no paved road network. Therefore, road proximity will be of little significance when considering surface irrigation suitability in these areas. The top ranking of elevation and slope gradients is expected since they collectively contribute the bulk of the establishment cost of surface irrigation through land preparation, labour cost, equipment installation as well as irrigation efficiency (Ali, 2010).

6.4.3 Potential surface irrigation suitable areas.

This study shows that the total surface irrigation suitable area in Nigeria is about 3.34 million hectares which is larger than previously reported estimates for Nigeria (FAO, 2012; You et al., 2011). However, the distribution of suitable areas across the entire country and the various regions compares well with previous reports (FAO, 2012). The higher suitability reported in this study than previous estimates could be due to the robustness of CI applied and more detailed analysis compared to previous studies which were mostly carried out at the continental scale with primary focus on the major river basins. The higher suitability of the northern region compared to the southern counterpart could be attributed to the relatively drier environments of the north as a result of harsh weather condition and the relatively flat landscape. Furthermore, the northern region hosts a larger portion of the major river basins in Nigeria (FAO, 2012).

The narrow distribution of CI suitability index around the mean indicates the robustness of CI to extreme values of decision criteria and the weights associated with them (Lee and Hu, 2013; Krishnan et al., 2015**)**. CI has the capacity to capture not only the importance but also the interaction among evaluation criteria in the decision aggregation process (Grabisch et al., 2008; Chakan et al., 2012). Therefore, the overall suitability index cannot be disproportionately influenced by any individual criterion. Fitting a generalized fuzzy membership function across the study area may have had an undue negative effect on irrigation suitability of some areas, especially around the Jos and Mambilla plateaus. Although these areas are on relatively high altitude, irrigation activity is still possible on some relatively flat spots at the peak of the plateau. Fitting a local membership function may have captured well the local variations across the study area.

Furthermore, the location of most current irrigation sites in moderate to marginally suitable areas is a reflection of the absence of quantitative approach to previous irrigation suitability assessments. Also, the absence of quantitative soil information in previous suitatbility assessment projects was obvious in this study as the majority of existing irrigation dams sited on marginally to non-suitable areas are within areas known for high sand content, especially around the Hadejia Jama'are river basin. The location of Bokkos irrigation dam is obviously on marginal to non-suitable area. This could be attributed to the limitation posed by the steep slope gradient and high elevation of the plateau as mentioned earlier. In sharp contrasts to the aforementioned irrigation locations, the Benue and Niger River basins depict highly suitable areas for surface water irrigation and presents a great opportunity for future irrigation projects in the country. If properly harness these two river basins have the capacity to serve as food hub of Nigeria.

Considering that over 70% of the total cropped area of Nigeria is in the northern region with low rainfall amount marked with shorter growing season, and rapid population growth, irrigation is an essential factor in any food security strategy for the country. Nigeria has a very wide range of agro-ecology that could support the growing of diverse crop. Therefore, with irrigation and proper crop management practices, farmers in this country can engage in all-year round farming activities for sustainable production of staples food such as rice, millet, maize as well as vegetables and cash crops.

6.5 Conclusion

Given the global water demand accentuated by adverse impact of climate change, assessment of irrigation potential is of prime importance for national planning towards sustainable agricultural production especially areas with unfavourable climatic conditions. In this study we have successfully applied Choquet integral function to irrigation suitability criteria and model the suitability index. Overall, about 3.34×10^6 ha of land in Nigeria is potentially suitable for surface water irrigation development. Out of the total suitable areas for surface water irrigation in Nigeria, only about 1% is suitable for large scale irrigation while about 15% and 84% is suitable for medium and small scale irrigation respectively. Major physical limitations are due to topography, slope and soil properties. This study demonstrates that the combination of the evaluation criteria by Choquet integral function and the modelling of their interaction degrees by fuzzy measures improved aggregation outputs. This study further confirms that Choquet fuzzy integral is a credible and robust approach for the multi-criteria assessment of data from different domains and sources to delineate areas that are suitable for surface water irrigation. The outputs of this study will be useful to researchers and policy makers at national and regional levels for projects aiming at sustainable agricultural development in Nigeria.

References

Akpa, S. I.C, Odeh, I. O. A, Bishop, T. F.A, & Hartemink, A. E. 2014. Digital mapping of soil particle-size fractions for Nigeria. *Soil Science Society of America Journal*, *78*(6), 1953-1966.

Akpa, S. I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E. & Amapu, I.Y. & Leenaars, J.G.B. 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma* (under review).

Ali, M. H. 2010. Fundamentals of irrigation and on-farm water management. Volume 1, Springer, New York 560p.

Altchenko, Y., & Villholth, K. G. 2014. Mapping irrigation potential from renewable groundwater in Africa-a quantitative hydrological approach. *Hydrology and Earth System Sciences Discussions*, *11*, 6065-6097.

Ayoola, J.B. 2009. National medium term priority framework for food security and agricultural development in Nigeria. FAO, Rome,Italy.

Baczynski, M. & Jayaram, B. 2008. Fuzzy Implications. Springer, 310 p.

Braimoh, A. K., Vlek, P. L. G., & Stein, A. 2004. Land evaluation for maize based on fuzzy set and interpolation. *Environmental Management*, 33(2), 226–238.

Burrough, P.A., MacMillan, R.A., van Deursen, W.P.A. 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography. *Journal of Soil Science*, 43, 193-210.

Chakan, A.A., Sarmadian, F., Heydari, A., Omid, M., Mohammadi, J., & Odeh, I. 2012. Performance assessment of fuzzy integrals as aggregation operators for soil suitability analysis (Case Study: Silakhor Plain of Lorestan Province). *Journal of Range and Watershed Management,* 65 (3), 301-314.

Chen, Y., Yu, J., & Khan, S. 2010. Spatial sensitivity analysis of multi-criteria weights in GIS-based land suitability evaluation. *Environmental Modelling & Software*, *25*(12), 1582-1591.

Chen, Y., Yu, J., & Khan, S. 2013. The spatial framework for weight sensitivity analysis in AHP-based multi-criteria decision making. *Environmental Modelling & Software*, *48*, 129-140.

Choquet, G. 1954. Theory of capacities. *Annales de l'institut Fourier* 5, 131-295.

Doumpos, M., & Zopounidis, C. 2011. A multicriteria outranking modeling approach for credit rating. *Decision Sciences*, *42*(3), 721-742.

FAO, 2014. WFP. The State of Food Insecurity in the World 2014 strengthening the enabling environment for food security and nutrition. FAO, Rome, Italy.

FAO, 2012. AQUASTAT, FAO's global information system on water and agriculture. http://www.fao.org/nr/aquastat.

FAO, 2009. Report of the Food and Agriculture Organisation expert meeting on how to feed the world in 2050, June 24–26. Rome, Italy.

Feizizadeh, B., & Blaschke, T. 2013. Land suitability analysis for Tabriz County, Iran: a multi-criteria evaluation approach using GIS. *Journal of Environmental Planning and Management,* 56:1, 1-23, DOI: 10.1080/09640568.2011.646964.

Frenken, K., & Faurès, J. M. 1997. *Irrigation potential in Africa: A basin approach* (Vol. 4). Food and Agriculture Organisation, Rome, Italy. 177p.

Frenken, K. (2005). *Irrigation in Africa in figures: AQUASTAT survey, 2005* (Vol. 29). Food and Agriculture Organisation, Rome, Italy. 74p.

Grabisch, M., Kojadinovic, I. & Meyer, P. 2008. A review of methods for capacity identification in Choquet integral based multi-attribute utility theory: Applications of the Kappalab R package. *European Journal of Operational Research* 186 (2), 766-785.

Guillaume, R., Houé, R., & Grabot, B. 2014. Robust competence assessment for job assignment. *European Journal of Operational Research*, *238*(2), 630-644.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*. 25, 1965–1978.

Joerin, F., Thériault, M., & Musy, A. 2001. Using GIS and outranking multicriteria analysis for land-use suitability assessment. *International Journal of Geographical Information Science*, *15*(2), 153-174.

Joss, B. N., Hall, R. J., Sidders, D. M., & Keddy, T. J. 2008. Fuzzy-logic modeling of land suitability for hybrid poplar across the Prairie Provinces of Canada. *Environmental Monitoring and Assessment*, *141*(1-3), 79-96.

Keshavarzi, A., Sarmadian, F., Heidari, A., & Omid, M. 2010. Land suitability evaluation using fuzzy continuous classification (a case study: Ziaran region). *Modern Applied Science*, *4*(7), p72.

Klement, E. P., & Mesiar, R. 2005. Logical, algebraic, analytic and probabilistic aspects of triangular norms. Elsevier. 492p.

Koczkodaj, W. W., Herman, M. W., and Orlowski, M. 1997 Using consistency-driven pairwise comparisons in knowledge-based systems. Proceedings of the sixth international conference on information and knowledge management, ACM Press, 91-96.

Kordi, M., & Brandt, S. A. 2012. Effects of increasing fuzziness on analytic hierarchy process for spatial multicriteria decision analysis. *Computers, Environment and Urban Systems*, 36(1), 43–53.

Kopecký M, Čížková Š. 2010. Using topographic wetness index in vegetation ecology: does the algorithm matter? *Applied Vegetation Science*, 13, 450–459. *DOI: 10.1111/j.1654-109X.2010.01083.x.*

Krishnan, A. R., Engku Abu Bakar, E. M. N., & Mat Kasim, M. 2015. A Hybrid Multiattribute Decision Making Model for Evaluating Students' Satisfaction towards Hostels. *Advances in Decision Sciences* (in press).

Kumbhar, V., Choudhury, S., Sen, A., & Singh, T. P. 2014. Assessment of Irrigation and Agriculture Potential Using Geospatial Techniques: A Case Study of "Bhima-Ujjani" Project. *Procedia-Social and Behavioral Sciences*, *157*, 277-284.

Lanni, C., McDonnell, J. J., & Rigon, R. 2011. On the relative role of upslope and downslope topography for describing water flow path and storage dynamics: a theoretical analysis. *Hydrological Processes*, *25*(25), 3909-3923.

Lauenroth, W. K., & Bradford, J. B. 2006. Ecohydrology and the partitioning AET between transpiration and evaporation in a semiarid steppe. *Ecosystems*, *9*(5), 756-767.

Lauenroth, W. K., & Bradford, J. B. 2009. Ecohydrology of dry regions of the United States: precipitation pulses and intraseasonal drought. *Ecohydrology*, *2*(2), 173-181.

Lee, S. K., & Yu, J. H. 2013. Composite indicator development using utility function and fuzzy theory. *Journal of the Operational Research Society*, *64*(8), 1279-1290.

Lust, T., & Rolland, A. 2013. Choquet optimal set in biobjective combinatorial optimization. *Computers & Operations Research*, *40*(10), 2260-2269.

Malczewski, J. 2006. Ordered weighted averaging with fuzzy quantifiers: GIS based multicriteria evaluation for land-use suitability analysis. *International Journal of Applied Earth Observation and Geoinformation*, 8(4), 270–277.

McBratney, A. B., & Odeh, I. O. 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, *77* (2), 85-113.

Minasny, B., & Hartemink, A. E. 2011. Predicting soil properties in the tropics. *Earth-Science Reviews*, *106*(1), 52-62.

Minasny B. & McBratney A.B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computational Geosciences*, 32, 1378–1388.

Mosadeghi, R., Warnken, J., Tomlinson, R., & Mirfenderesk, H. 2013. Uncertainty analysis in the application of multi-criteria decision-making methods in Australian strategic environ-mental decisions. *Journal of Environmental Planning and Management,* 56(8), 1097–1124.

Mosadeghi, R., Warnken, J., Tomlinson, R., & Mirfenderesk, H. 2015. Comparison of Fuzzy-AHP and AHP in a spatial multi-criteria decision making model for urban land-use planning. *Computers, Environment and Urban Systems*, *49*, 54-65.

Murofushi, T., & Sugeno, M. 1991. Fuzzy t-conorm integral with respect to fuzzy measures: generalization of Sugeno integral and Choquet integral. *Fuzzy Sets and Systems*, *42*(1), 57-71.

Mustafa, A. A., Singh, M., Sahoo, R. N., Ahmed, N., Khanna, M., Sarangi, A., & Mishra, A. K. 2011. Land suitability analysis for different crops: a multi-criteria decision making approach using remote sensing and GIS. New Delhi-110 012: *Indian Agricultural Research Institute, New Delhi*.

Odeh, I. & Crawford, M. 2009. On fuzzy measures, choquet integral and classical aggregation operators for multi-dimensional decision-making in soil suitability assessment. *Biennial Meeting of Commission 1.5 Pedometrics Division 1 of the International Union of Soil Science (IUSS) held* August 26-28. China Agricultural University Beijing, China.

Peneva, V. & Popchev, I. 2003. Properties of the aggregation operators related with fuzzy relations. *Fuzzy Sets and Systems* 139, 615-633.

R Development Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ricker-Gilbert, J., Jumbe, C., & Chamberlin, J. 2014. How does population density influence agricultural intensification and productivity? Evidence from Malawi. *Food Policy*, *48*, 114-128.

Rowley, H. V., Geschke, A., & Lenzen, M. 2015. A practical approach for estimating weights of interacting criteria from profile sets. *Fuzzy Sets and Systems*, *272*, 70-88.

Roy, B. 2010. Robustness in operational research and decision aiding: A multi-faceted issue. *European Journal of Operational Research*, *200*(3), 629-638.

Saaty, T. L. 2008. “Decision Making with the Analytic Hierarchy Process”, *International Journal of Services Sciences*, 1(1), 83-98.

Saaty, T. L & Vargas, L.G. 2001. Models, methods, concepts & applications of analytical hierarchy process. International Series in Operations Research and Management sciences. London, Kluwer Academic Publishers.

Shapley, L. S. 1953. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, *39*(10), 1095.

Sicat, R. S., Carranza, E. J. M., & Nidumolu, U. B. 2005. Fuzzy modeling of farmers' knowledge for land suitability classification. *Agricultural Systems*, *83*(1), 49-75.

Sousa, J.M.C. & Kaymak, U. 2002. Fuzzy Decision Making, Modeling and Control, World Scientific Publishing Company, 335p.

Sugeno, M. 1974. Theory of fuzzy integrals and its application, Doctoral Thesis, Tokyo Institute of Technology.

Takahagi, E. 2005a. Usage: Fuzzy Measure-Choquet Integral Calculation System (λ fuzzy measure and sensitivity analysis). Accessed on 02/02/2014 http://www.isc.senshu-u.ac.jp/~thc0456/Efuzzyweb/mant2/mant2.html.

Takahagi, E. 2005b. λ fuzzy measure identification methods using λ and weights. *Senshu University, Japan*.

Takeshima, H. & Adesugba, M. 2014. Irrigation Potential in Nigeria: Some Perspectives Based on Factor Endowments, Tropical Nature, and Patterns in Favorable Areas. IFPRI Discussion Paper 01399.

Thornton, P. K., Jones, P. G., Alagarswamy, G., & Andresen, J. 2009. Spatial variation of crop yield response to climate change in East Africa. *Global*

Environmental Change, *19*(1), 54-65.

Torra, V. 2003. Twofold integral: a Choquet integral and Sugeno integral generalization, *Butlletı´ de l'Associacio´ Catalana de'Intellige`ncia Artificial*, 29,13–19 (in Catalan), Preliminary version: IIIA Research Report TR-2003-08 (in English).

Torra, V. & Narukawa, Y. 2006. The interpretation of fuzzy integrals and their application to fuzzy systems. Int. J. Approximate Reasoning 41, 43-58.

Trabucco, A., and Zomer, R.J. 2009. *Global Aridity Index (Global-Aridity) and Global Potential Evapo-Transpiration (Global-PET) Geospatial Database*. CGIAR Consortium for Spatial Information. Published online, available from the CGIAR-CSI GeoPortal at: http://www.csi.cgiar.org.Accessed 12/06/14.

Triantafilis, J., Ward, W. T., & McBratney, A. B. 2001. Land suitability assessment in the Namoi Valley of Australia, using a continuous model. *Soil Research*, *39*(2), 273-289.

United Nations (UN), 2011. World Population Prospects: The 2010 Revision. New York, New York, USA.

United Nations Environment Programme (UNEP), 1997. World atlas of desertification 2nd edn. UNEP, London.

Worqlul, A. W., Collick, A. S., Rossiter, D. G., Langan, S., & Steenhuis, T. S. 2015. Assessment of surface water irrigation potential in the Ethiopian highlands: The Lake Tana Basin. *Catena*, *129*, 76-85.

Yang, R., Wang, Z., Heng, P.N. & Leung, K.S. 2005. Fuzzy numbers and fuzzification of the Choquet integral. Fuzzy Sets and Systems 153, 95-113.

Zhang, K., & Achari, G. 2010. Uncertainty propagation in environmental decision making using random sets. Procedia Environmental Sciences, 2:576–584.

Zimmermann, H.J. 1992. Fuzzy Set Theory and its Applications, 2nd Edition, Kluwer Academic press, 399p.

Chapter 7.

General discussions, conclusions and future work

7.1 General discussions

Globally there is a rising need of spatially explicit soil information to support targeted and specific decisions on global and regional challenges posed by climate change, food and water shortage, land degradation, and loss of biodiversity (Arrouays et al., 2014; Grunwald et al., 2015). This is of particular importance in most developing countries, especially in sub-Saharan Africa which is currently plagued by poverty, hunger and land degradation. In the light of the scarcity of quantitative soil information in SSA, there is need to optimize the use of legacy soil data to meet modern soil information needs. Before now, there is it not much work done in this regards in the SSA, especially at the national scale.

This thesis demonstrates how sparse legacy soil data could be used to populate national geodatabase with relevant soil information that will support national and regional planning. In Chapter 1 the general background of the thesis and outline of each research chapters were presented. In Chapter 2 a review of the need for digital soil mapping and soil information in developing countries highlighting on how best to utilize sparse legacy soil data to deliver soil information that are relevant for national scale planning. Further reviews were on the various techniques at the disposal of digital soil mapping scientists, especially those amenable to complex and sparse data condition. From the review it was clear that data mining models (because of their capacity to exploit complex data structure) are most robust for scarce data and complex conditions. However, not much has been done previously to optimize the enormous potentials in these modelling tools for DSM in data-sparse region, particulary in the SSA. It was also revealed that there has not been much DSM work on the prediction of PSFs as compositional data and none of the few reported works in this regard predicted PSFs beyond the soil surface depth (Odeh et al., 2003; Buchanan et al., 2012; Niang et al., 2014). Another research gap found from the review was the fact that not much has been done to add extra value to DSM primary products to support national and regional scale planning.

In addressing one of the research gaps identified in Chapter 2, the focus of Chapter 3 was to predict PSFs as a compositional data at national scale. In doing this the robustness of Random Forest model (RFM) was tested in comparison with Cubist models and the popular multiple linear regression (MLR). First, mass preserving spline functions, which has exhibited the capacity to estimate soil properties at predetermined soil depth were fitted to PSFs data to output values at the six standard depth intervals (Arrouays et al., 2014). Further, additive log-ratio (ALR) transformation technique (Aitchison et al., 2000) was employed prior to model prediction to ensure that predicted PSFs some up to a constant value of 100 as expected. Random forest model (RFM) turns out to predict PSFs in the study area better than MLR and Cubist. Inclusion of sampling depth as a predictor substantially improved prediction accuracy of RFM, especially at the lower depth intervals.

The predicted PSFs are useful inputs in hydrological models for soil erosion and climate modellings, as well as decision criteria in various soil and land suitability assessments as demonstrated in Chapter 6. Another important use of the predicted PSFs is in the development and/or calibration of pedotransfer functions to estimate difficult-to-measure soil properties such as soil hydraulic (Rajkai et al., 1996; Arya et al., 1999; Wagner et al., 2001), soil moisture retention capacity (Botula et al., 2012; 2013), CEC and bulk density as demonstrated in Chapter 5 of this thesis. All these derivatives of PSFs will be vital input in national soil and environmental monitoring programmes and to guide decisions on environmental risk managements.

The need for additional sequestration of carbon (C) in the terrestrial agro-ecosystems has dominated discussions among scientists and policy makers around the globe in recent time. While several studies have shown that land use change can affect carbon content and sequestration of soils, the magnitude and dynamics of these changes in different ecosystems especially in the SSA have not been extensively studied. Chapter 4 bordered on the estimation of total SOC and carbon sequestration of soils in Nigeria and their variation across and within different agro-ecological zones (AEZ). The soil-landscape modelling approach was used to estimate SOC and carbon stock for the entire country. Therafter, the mean SOC density of the different land use/land cover (LULC) types across the various AEZ were estimated. These were then used to calculate the difference between the prevailing LULC types and the pristine LULC types in each zone. The assumption was that the difference between the mean SOC density of any LULC type and the pristine LULC in a particular AEZ is the amount of SOC that soils could sequester by restorating the target land use. Results indicate that soils in the Derived Savannah (DS) and Sahel Savannah (SHS) show the greatest capacity to sequester additional C while about 6.5 Pg C with an average density of 71.60 Mg C ha^{-1} abound in the top 1 m of soil depth of the entire Nigeria. Restoration of the various landuse types to their natural ecosystem, has the potential to sequester about 0.2 to 30.8 Mg C ha^{-1} depending on the LULC and AEZ.

Although one may favourably argue that the assumption used to estimate carbon sequestration potential in this study may not be feasible considering the high food demand and the need to put more land into cultivation. However, if only a faction of the potential carbon sequestration is attained it will go a very long way in the current fight against land degreadtion and global warming. Furthermore, this kind of national scale study can elucidate hotspots for future carbon accounting or land use restoration programmes. Another strong argument against the output of this particular research may be the varying age of the legacy SOC data used in estimating the SOC stocks. Notwithtanding, the estimated total SOC stock is still useful and could form a baseline for future SOC studies (Bui et al., 2009) or input in future ecosystem monitoring programmes in the country. The sparse nature of the SOC dataset used in this study conferred the wide prediction interval observed for the estimated SOC stocks. This may also impact negatively on the reliability of this output. High uncertainty in SOC stock prediction has been reported for studies in the Limpopo park of Mozambique (Cambule et al., 2014).

As disussed earlier in Chapter 2, legacy soil data are usually characterized by inconsistency as well as incompleteness and as such it is not uncommon to see most difficult to measure but very important soil attributes missing in national soil database. The use of pedotransfer functions has been salvaging this issue especially at the field or catchment scale for several decades now. However, despite the enormous gains in the use of pedotransfer functions to remedying the problem of incomplete soil database, the application of most PTFs in national scale studies especially in developing countries is hampered by the amount and structure of input data. Chapter 5 covered the use of different data grouping techniques and combination of soil and environmental attributes to enhance the performance of PTFs for national-scale studies in scarce data condition. Subdividing the input data into different groups based on soil depth and incorporating environmental data such as climate, topography and vegetation attributes resulted in more accurate BD predictions. However, results did not show any advantage of combining soil and environmental data for ECEC prediction. The reasonable prediction accuracy from bulk density PTFs using only particle-size fractions in this study could reduce the extra cost from the use SOC data in existing PTFs for bulk density. Also the findings of this research will help relief modellers and hydologists the burden of acquiring soil bulk density for large scale studies.

Agriculture intensification through irrigation expansion has been proffered as a sure way out of the lingering food crisis in SSA and most developing countries, most especially with the unpredictable and erratic nature of rainfall in most of these countries. Therefore, in the context of digital soil assessment (Carré et al., 2007), additional steps were taken in Chapter 6 to add values to some of the primary DSM products obtained in the previous Chapters. This is to make them more relevant for national and regional developmental planning. Here, Choquet fuzzy integral (CI) aggregation technique was employed in mapping suitability for surface irrigation in Nigeria. This was achieved through multi-criteria assessment of potential evaluation criteria comprising of some of the soil and environmental attributes obtained in Chapter 3 and some selected socio-economic variables. Results indicate that CI is a better aggregation operator than the classical weighted mean operator. This is in line with previous studies that have shown that fuzzy set theory is powerful and very flexible in dealing with the complexities and uncertainties embedded in land suitability assessment (Braimoh et al., 2004; Sicat et al., 2005; Joss et al., 2008; Odeh and Crawford, 2009; Chakan et al., 2012).

About 3.4 million hectares of land out of the total landmass of Nigeria is suitable for irrigation, while the northern region with over 80% suitability is the most viable hotspot for irrigation development in the country. Furthermore, substantial potential of small-scale irrigation expansion exists in Nigeria for dry-season/high value crop production. Output of this research will form an integral part of decisions making for cultivation of specialized crops such as wheat, cotton, sugarcane, rice among other high water demanding crops. Well planned irrigation project will not only boost the food bank of most countries in SSA but will also increase the income and of the populace since agriuculture is the highest employer of labour in the SSA.

7.2 General conclusions

This study provides an example of how a geodatabase of important soil attributes can be populated from a limited soil data set. All potential models are as good as the quality of the input datasets. Therefore, outputs of this research are good first approximations of digital mapping of soil attributes under the sparse data condition. No doubt, there is a need to continue to improve on these first approximations as more data becomes available. From the studies presented in this thesis, the following salient points could be highlighted:

1. DSM is an efficient but challenging quantitative spatial prediction approach, especially in a data-scarce situation. We have demonstrated the robustness of RFM to predict soil functional properties in such condition.
2. The combination of environmental attributes with soil properties is a sure way to developing PTFs for a national scale and under sparse legacy data data condition.
3. Derived savannah and Sahel savannah agroeclogical zones in Nigeria are potential hotspots for any future carbon accounting or ecosystem monitoring programme in the country.
4. Choquet integral is effective aggregator in irrigation suitability assessment using multi-criteria approach. About 3.4 million hactares of land is potentially suitable for irrigation agriculture in Nigeria. This presents a viable option to increased agricultural production to meet the food demands of the growing population of Nigeria.

7.3 Future research

1. In this work, some gaps in data coverage and spread in Nigeria were identified. Although there is currently no hope of any new national soil survey project in the

nearest future, additional soil profile data could be sourced from private surveys by researchers and students of the various educational and research institutions in Nigeria.

2. The uncertainties associated with prediction models used in this study were exemplified for SOC. However, there is a need for an extensive uncertainty analysis for all other predicted soil properties and to propagate these uncertainties through subsequent products derived from the primary soil attributes. There is also the need to evaluate other sources of uncertainty apart from the model uncertainty as demonstrated by Nelson et al. (2011).

3. In this study, few PTFs were developed for soil bulk density and cation exchange capacity with an enhancement to suit data scarce situations. There is a need for further development of PTFs to estimate other key functional properties like hydraulic conductivity, available water capacity, phosphorus retention capacity and soil erodibility index, among others.

4. Although the soil data used for the PTFs developed in this study were from Nigeria. It will be nice to evaluate the transferability of the developed PTFs to other neighbouring Africa countries with peculiar soil data problems and related soil conditions. In the same way, it will also be nice to test the applicability of PTFs developed in other countries within similar soil domain as Nigeria, especially for those functional properties not covered in this thesis.

5. Following from point number 4 above, there is need to demonstrate the robustness of the multi-criteria approach to irrigation suitability analysis employed in this study in smaller scale studies such as at the catchment or field level and perhaps to other countries in the SSA. Also, the concept of crop versatility which encompasses the productivity of various crops need to be incorporated into the irrigation suitability assessment especially for economic or cash crops like cocoa,

rice, palm tree, sugar cane, etc. In addition there is need to consider the economic viability of agricultural intensification through irrigation development. One way to go about this will be to incorporate hydrological models like SWAT to quantify the actual water available for irrigation vis-à-vis the potential irrigation suitability. The net economic gain from possible adoption of irrigation agriculture could also be quantified using the economic models by incorporating crop yield and market price data.

References

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonça Santos, M.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A. & Zhang, G.L. 2014. Chapter Three-GlobalSoilMap: toward a fine-resolution global grid of soil properties. *Advances in Agronomy*. 125, 93–134.

Arya, L. M., Leij, F. J., Shouse, P. J., & van Genuchten, M. T. 1999. Relationship between the hydraulic conductivity function and the particle-size distribution. *Soil Science Society of America Journal*, *63*(5), 1063-1070.

Botula, Y. D., Cornelis, W. M., Baert, G., & Van Ranst, E. 2012. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (DR Congo). *Agricultural Water Management*, *111*, 1-10.

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).

Braimoh, A. K., Vlek, P. L. G., & Stein, A. 2004. Land evaluation for maize based on fuzzy set and interpolation. *Environmental Management*, 33(2), 226–238.

Buchanan, S., Triantafilis, J., Odeh, I., & Subansinghe, R. 2012. Digital soil mapping of compositional particle-size fractions using proximal and remotely sensed ancillary data. *Geophysics*, 77(4), 201-211.

Bui, E., Henderson, B., & Viergever, K. 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. *Global Biogeochemical Cycles*, *23*(4).

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.

Carré, F., McBratney, A. B., Mayr, T., & Montanarella, L. 2007. Digital soil assessments: Beyond DSM. *Geoderma*, *142*(1), 69-79.

Chakan, A.A., Sarmadian, F., Heydari, A., Omid, M., Mohammadi, J., & Odeh, I. 2012. Performance assessment of fuzzy integrals as aggregation operators for soil suitability analysis (Case Study: Silakhor Plain of Lorestan Province). *Journal of Range and Watershed Management*, 65 (3), 301-314.

Grealish, G. J., Fitzpatrick, R. W., & Hutson, J. L., 2015. Soil survey data rescued by means of user friendly soil identification keys and toposequence models to deliver soil information for improved land management. *Geographical Research Journal*, *6*, 81-91.

Joss, B. N., Hall, R. J., Sidders, D. M., & Keddy, T. J. 2008. Fuzzy-logic modeling of land suitability for hybrid poplar across the Prairie Provinces of Canada. *Environmental Monitoring and Assessment*, *141*(1-3), 79-96.

Nelson, M. A., Bishop, T. F. A., Triantafilis, J., & Odeh, I. O. A. 2011. An error budget for different sources of error in digital soil mapping. *European Journal of Soil Science*, *62*(3), 417-430.

Niang, M. A., Nolin, M. C., Jégo, G., & Perron, I. 2014. Digital Mapping of Soil Texture Using RADARSAT-2 Polarimetric Synthetic Aperture Radar Data. *Soil Science Society of America Journal*, *78*(2), 673-684.

Odeh, I. O., Todd, A. J., & Triantafilis, J. 2003. Spatial prediction of soil particle-size fractions as compositional data. *Soil Science*, *168*(7), 501-515.

Rajkai, K., Kabos, S., Van Genuchten, M. T., & Jansson, P. E. 1996. Estimation of water-retention characteristics from the bulk density and particle-size distribution of swedish soils. *Soil Science*, *161*(12), 832-845.

Wagner, B., Tarnawski, V. R., Hennings, V., Müller, U., Wessolek, G., & Plagge, R. 2001. Evaluation of pedo-transfer functions for unsaturated soil hydraulic conductivity using an independent data set. *Geoderma*, *102*(3), 275-297.

Appendix

Appendix 3.1 Performance of RFM, Cubist and MLR in modelling Particle-size fractions

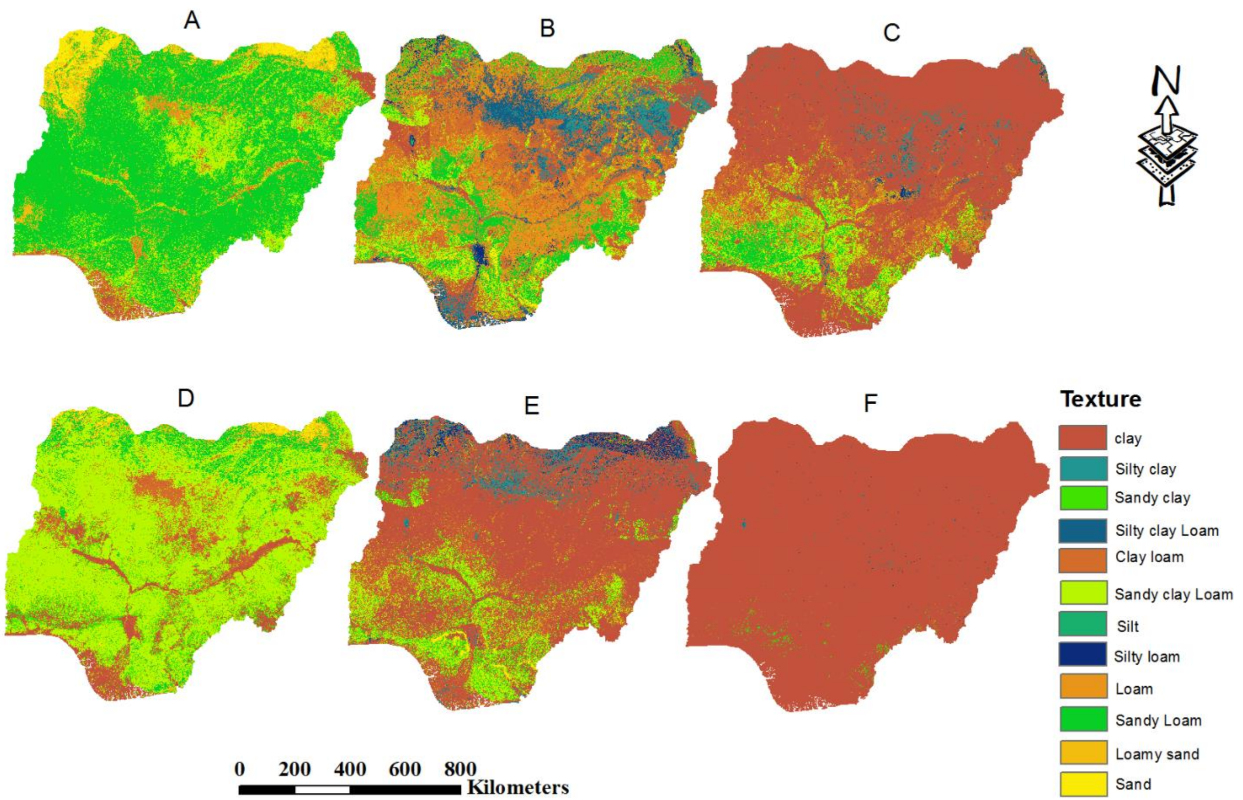
		Random Forest				Cubist				MLR			
PSF	Depth	ME	RMSE	R^2	p_c	ME	RMSE	R^2	p_c	ME	RMSE	R^2	p_c
Clay	0-5	3.53	13.59	0.53	0.65	2.52	13.98	0.48	0.66	3.07	14.44	0.45	0.63
	5-15	3.40	13.11	0.56	0.69	2.23	13.35	0.53	0.70	2.97	14.64	0.44	0.63
	15-30	2.95	13.38	0.54	0.68	1.28	15.61	0.39	0.62	2.45	14.93	0.42	0.62
	30-60	2.60	14.98	0.42	0.62	2.18	16.32	0.35	0.57	3.34	15.99	0.35	0.55
	60-100	4.01	15.70	0.29	0.46	2.18	18.12	0.16	0.39	3.80	16.74	0.21	0.41
	100-200	4.21	15.60	0.16	0.30	2.18	18.71	0.04	0.19	3.69	16.5	0.07	0.21
Sand	0-5	-6.51	19.67	0.48	0.60	-4.36	19.69	0.44	0.63	-6.16	20.52	0.42	0.59
	5-15	-6.03	19.26	0.49	0.63	-4.33	19.43	0.47	0.66	-5.96	20.33	0.43	0.60
	15-30	-5.26	18.79	0.49	0.63	-2.73	20.72	0.39	0.62	-4.95	19.78	0.44	0.62
	30-60	-4.14	18.81	0.43	0.61	-2.96	16.32	0.35	0.58	-4.98	19.5	0.41	0.59
	60-100	-5.71	19.48	0.33	0.50	-3.00	21.55	0.21	0.44	-6.02	20.09	0.30	0.47
	100-200	-6.67	19.86	0.21	0.36	-4.37	21.89	0.08	0.24	-6.39	20.65	0.14	0.29
Silt	0-5	2.99	12.22	0.26	0.39	1.84	12.5	0.23	0.43	3.09	13.79	0.10	0.26
	5-15	2.63	11.72	0.27	0.42	2.10	12.03	0.23	0.42	2.98	13.22	0.11	0.26
	15-30	2.31	10.96	0.25	0.39	1.44	11.33	0.22	0.44	2.50	12.1	0.12	0.28
	30-60	1.54	9.82	0.24	0.41	0.78	10.59	0.19	0.42	1.64	10.49	0.15	0.32
	60-100	1.69	9.73	0.24	0.40	0.82	9.83	0.24	0.44	2.17	16.74	0.12	0.26
	100-200	2.46	10.06	0.21	0.35	2.19	9.92	0.22	0.36	2.70	11.03	0.09	0.22

†PSF; Particle size fractions, ME; Mean error, RMSE; Root mean square error, p_c, Lin's concordance correlation coefficient

Appendix 3.2 Performance of RMF, Cubist and MLR in modelling particle-size fractions with the inclusion of soil depth as a predictor

		Random Forest				Cubist				MLR			
PSF	Depth	ME	RMSE	R^2	p_c	ME	RMSE	R^2	p_c	ME	RMSE	R^2	p_c
Clay	0-5	-0.40	6.48	0.89	0.94	-0.32	5.92	0.91	0.95	0.15	15.23	0.41	0.56
	5-15	0.35	5.71	0.91	0.95	0.01	5.47	0.91	0.96	0.66	13.2	0.52	0.62
	15-30	2.26	6.93	0.89	0.94	1.33	6.05	0.91	0.95	3.04	14.98	0.45	0.57
	30-60	-0.23	10.04	0.72	0.85	1.29	9.35	0.77	0.87	6.25	17.16	0.3	0.46
	60-100	1.58	10.56	0.68	0.82	2.31	11.24	0.66	0.81	6.10	17.58	0.25	0.45
	100-200	1.18	12.63	0.43	0.64	-0.66	13.47	0.41	0.63	-1.89	19.01	0.04	0.2
Sand	0-5	0.03	7.70	0.91	0.95	0.10	6.15	0.94	0.97	-2.05	19.97	0.39	0.58
	5-15	-0.42	7.05	0.92	0.96	-0.12	6.38	0.94	0.97	-2.00	18.77	0.45	0.6
	15-30	-1.97	7.26	0.92	0.96	-1.21	6.70	0.93	0.96	-2.85	19.81	0.38	0.57
	30-60	-0.74	9.55	0.85	0.92	-1.83	9.35	0.85	0.92	-8.45	21.25	0.37	0.52
	60-100	-2.09	11.52	0.76	0.87	-2.41	12.12	0.76	0.87	-7.46	21.82	0.28	0.48
	100-200	-1.91	15.85	0.51	0.70	0.27	15.51	0.55	0.74	-0.90	24.52	0.05	0.22
Silt	0-5	0.37	4.44	0.88	0.94	0.22	3.45	0.93	0.96	1.90	12.33	0.15	0.33
	5-15	0.08	4.14	0.90	0.95	0.11	3.53	0.93	0.96	1.34	12.33	0.15	0.32
	15-30	-0.29	3.43	0.91	0.95	-0.12	2.97	0.94	0.97	-0.19	11.95	0.07	0.23
	30-60	0.97	5.17	0.82	0.89	0.54	4.64	0.85	0.92	2.20	11.27	0.17	0.33
	60-100	0.51	5.66	0.76	0.85	0.10	5.19	0.79	0.89	1.37	17.58	0.14	0.31
	100-200	0.74	7.15	0.59	0.74	0.39	6.80	0.63	0.79	2.79	11.26	0.07	0.19

†PSF; Particle size fractions, ME; Mean error, RMSE; Root mean square error, p_c, Lin's concordance correlation coefficient



Appendix 3.3 Spatial distribution of predicted soil texture at the 5-15cm and 60-100 depth interval based on RFM (A & D), Cubist (B & E) and MLR (C & F).

Appendix 5.1. Pedotransfer functions for predicting Bulk density for Nigeria.

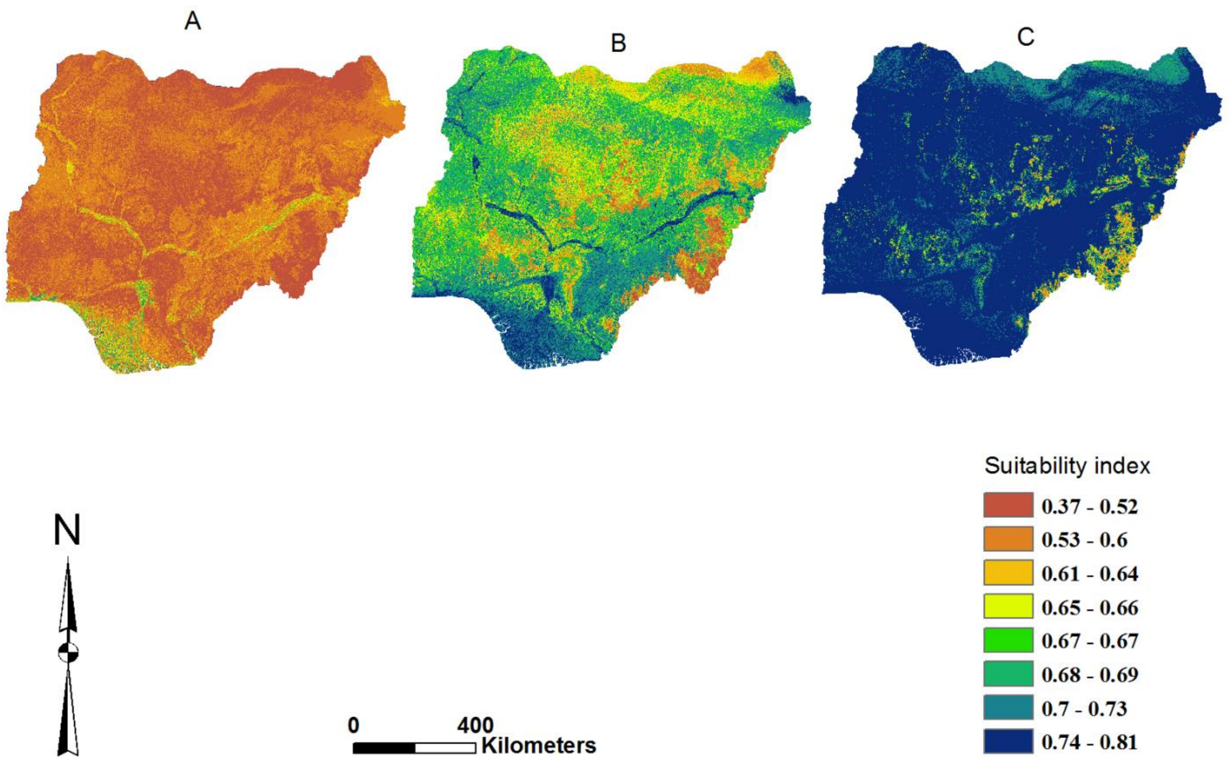
Model	Equation
	All Data
PTF-1	BD=1.177+0.00263Sand-0.0439logSilt+0.00208Silt
PTF-2	BD=0.903+0.00283-0.0958LogEL-0.00184SPI-1.355NDVI+1.451EVI-0.0251logFLACC+0.00853TWI-0.00014Asp
PTF-3	BD=1.440+0.000499Temp+0.00256Sand-0.0714logSL-0.112logEL-0.0174logFlACC-0.00011Asp-1.331NDVI+1.429EVI+3.330ProfC-0.00331SPI
	Topsoil
PTF-1	BD=1.172+0.00250Sand-0.0341logSilt+0.000877Silt
PTF-2	BD=0.885+0.0077TWI+0.00277Temp+0.0486logSL-0.0246logFLACC-0.0370logEL+1.444EVI-1.604NDVI
PTF-3	BD=-35.0793+0.00296Sand-1.271NDVI-4.517SPI+1.0774EVI-0.0346logEL-0.0174logSilt+4.507TWI+20.718logSL
	Subsoil
PTF-1	BD=1.512-0.00322Clay-0.0865logSilt
PTF-2	BD=-36.814-0.1340logEL+21.486logSL-0.0251logFLACC-0.00021Asp+1.523EVI-1.267NDVI+0.00261Temp+4.68TWI-4.674SPI
PTF-3	BD=-15.203+0.00322Sand-1.0301NDVI-2.0518SPI+0.817EVI-0.0195logEL-0.0191logSilt+2.0422TWI+9.347logSL

Asp, aspect; BD, Bulk density; EL, elevation; EVI, Enhanced vegetation index; FLACC, flow accumulation; NDVI, Normalized difference vegetation index; ProfC, profile curvature; Temp, Temperature; TWI, Topographic wetness index; SL; slope gradient; SPI, Stream power index.

Appendix 5.2. Pedotransfer functions for predicting Effective cation exchange capacity for Nigeria.

Model	Equation
	All Data
PTF-1	ECEC=-15.681+0.118Clay+4.097pH-0.124Sand+0.0887SOC
PTF-2	ECEC=19.231-77.282EVI-4.474logSL+36.553NDVI-0.00872Asp-0.0033FLACC-0.00176Rainfall+0.395TWI
PTF-3	ECEC=-12.130+0.223Clay+3.929pH-45.789EVI+0.157SOC-1.955logSL-0.0733Sand+21.436NDVI-0.00334EL
	Topsoil
PTF-1	ECEC=-22.612+0.212Clay+5.0295pH-0.110Sand+0.136SOC
PTF-2	ECEC=7.829-7.835logSL-84.369EVI+37.620+0.0412Temp-0.00257-0.164SPI
PTF-3	ECEC=-23.250+0.414Clay+4.614pH+0.0547Silt-0.520SPI-13.775EVI-0.0306SOC+0.141logSL-0.00794EL
	Subsoil
PTF-1	ECEC=-19.820+4.744pH-0.119Sand+0.137Clay+0.279SOC
PTF-2	ECEC=127.406-85.233EVI-5.942logSL+37.179NDVI-0.359Temp-0.0104Asp-0.00412Rainfall-0.00387FLACC+0.826SPI+145.637PlanC
PTF-3	ECEC=-24.453+0.413Clay+4.741pH+0.0557Silt-0.510SPI-11.873EVI-0.0365SOC+0.0885logSL-0.00858EL

Asp, aspect; ECEC, Effective cation exchange capacity; EL, elevation; EVI, Enhanced vegetation index; FLACC, flow accumulation; NDVI, Normalized difference vegetation index; PlanC, plan curvature; Temp, Temperature; SPI, Stream power index.



Appendix 6.1. Irrigation suitability maps based on different interaction indices (A; λ=0.35, B; λ=0.5, C; λ=0.65)