The application of deep learning for remote sensing of soil organic carbon stocks distribution in South Africa

Odebiri Omosalewa Olamide

217081218

A dissertation submitted in fulfilment for the Degree of Doctor of Philosophy in Environmental Sciences, in the School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal

Supervisor: Professor Onisimo Mutanga

Supervisor: Professor John Odindi

Pietermaritzburg, South Africa

JUNE 2022

Source: Jason Wheatley Duke Energy | illumination

Source: @boomplaatsorganicfarm

"Soil organic carbon is a valuable resource, but all soil organic carbon is not created equal"

Francesca Cotrufo & Jocelyn Lavallee

Abstract

Soil organic carbon (SOC) is a vital measure for ecosystem health and offers opportunities to understand carbon fluxes and associated implications. However, unprecedented anthropogenic disturbances have significantly altered SOC distribution across the globe, leading to considerable carbon losses. In addition, reliable SOC estimates, particularly over large spatial extents remain a major challenge due to among others limited sample points, quality of simulation data and suitable algorithms. Remote sensing (RS) approaches have emerged as a suitable alternative to field and laboratory SOC determination, especially at large spatial extent. Nevertheless, reliable determination of SOC distribution using RS data requires robust analytical approaches. Compared to linear and classical machine learning (ML) models, deep learning (DL) models offer a considerable improvement in data analysis due to their ability to extract more representative features and identify complex spatial patterns associated with big data. Hence, advancements in remote sensing, proliferation of big data, and deep learning architecture offer great potential for large-scale SOC mapping. However, there is paucity in literature on the application of DL-based remote sensing approaches for SOC prediction. To this end, this study is aimed at exploring DL-based approaches for the remote sensing of SOC stocks distribution across South Africa. The first objective sought to provide a synopsis of the use of traditional neural network (TNN) and DL-based remote sensing of SOC with emphasis on basic concepts, differences, similarities and limitations, while the second objective provided an in-depth review of the history, utility, challenges, and prospects of DL-based remote sensing approaches for mapping SOC. A quantitative evaluation between the use of TNN and DL frameworks was also conducted. Findings show that majority of published literature were conducted in the Northern Hemisphere while Africa have only four publications. Results also reveal that most studies adopted hyperspectral data, particularly spectrometers as compared to multispectral data. In comparison to DL (10%), TNN (90%) models were more commonly utilized in the literature; yet, DL models produced higher median accuracy (93%) than TNN (85%) models. The review concludes by highlighting future opportunities for retrieving SOC from remotely sensed data using DL frameworks.

The third objective compared the accuracy of DL—deep neural network (DNN) model and a TNN—artificial neural network (ANN), as well as other popular classical ML models that include random forest (RF) and support vector machine (SVM), for national scale SOC mapping using Sentinel-3 data. With a root mean square error (RMSE) of 10.35 t/ha, the DNN model produced the best results, followed by RF (11.2 t/ha), ANN (11.6 t/ha), and SVM (13.6

t/ha). The DNN's analytical abilities, combined with its capacity to handle large amounts of data is a key advantage over other classical ML models. Having established the superiority of DL models over TNN and other classical models, the fourth objective focused on investigating SOC stocks distribution across South Africa’s major land uses, using Deep Neural Networks (DNN) and Sentinel-3 satellite data. Findings show that grasslands contributed the most to overall SOC stocks (31.36 %), while urban vegetation contributed the least (0.04%). Results also show that commercial (46.06 t/h) and natural (44.34 t/h) forests had better carbon sequestration capacity than other classes. These findings provide an important guideline for managing SOC stocks in South Africa, useful in climate change mitigation by promoting sustainable land-use practices.

The fifth objective sought to determine the distribution of SOC within South Africa’s major biomes using remotely sensed-topo-climatic data and Concrete Autoencoder-Deep Neural Networks (CAE-DNN). Findings show that the CAE-DNN model (built from 26 selected variables) had the best accuracy of the DNNs examined, with an RMSE of 7.91 t/h. Soil organic carbon stock was also shown to be related to biome coverage, with the grassland (32.38%) and savanna (31.28%) biomes contributing the most to the overall SOC pool in South Africa. forests (44.12 t/h) and the Indian ocean coastal belt (43.05 t/h) biomes, despite having smaller footprints, have the highest SOC sequestration capacity. To increase SOC storage, it is recommended that degraded biomes be restored; however, a balance must be maintained between carbon sequestration capability, biodiversity health, and adequate provision of ecosystem services. The sixth objective sought to project the present SOC stocks in South Africa into the future (i.e. 2050). Soil organic carbon variations generated by projected climate change and land cover were mapped and analysed using a digital soil mapping (DSM) technique combined with space-for-time substitution (SFTS) procedures over South Africa through 2050. The potential SOC stocks variations across South Africa's major land uses were also assessed from current (2021) to future (2050). The first part of the study uses a Deep Neural Network (DNN) to estimate current SOC content (2021), while the second phase uses an average of five WorldClim General Circulation Models to project SOC to the future (2050) under four Shared Socio-economic Pathways (SSPs). Results show a general decline in projected future SOC stocks by 2050, ranging from 4.97 to 5.38 Pg, compared to estimated current stocks of 5.64 Pg. The findings are critical for government and policymakers in assessing the efficacy of current management systems in South Africa.

Overall, this study provides a cost-effective framework for national scale mapping of SOC stocks, which is the largest terrestrial carbon pool using advanced DL-based remote sensing approach. These findings are valuable for designing appropriate management strategies to promote carbon uptake, soil quality, and measuring terrestrial ecosystem responses and feedbacks to climate change. This study is also the first DL-based remote sensing of SOC stocks distribution in South Africa.

Keywords: Soil organic carbon; Remote sensing; Deep learning; Climate change; Sentinel 3; Land use; Land-use planning; Biomes, Climate; Topography; Management

Preface

This study was conducted in the School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa, under the supervision of Professor Onisimo Mutanga and Professor John Odindi from January 2020 to June 2022.

I declare that the current work represents my own ideas and has never been submitted to any other academic institutions. Acknowledgment has been duly made for statements originating from other authors.

Odebiri Omosalewa Olamide Signed [redacted] Date …06/06/2022…

1. Professor Onisimo Mutanga (Supervisor) Signed ………………. Date …06/06/2022….

2. Professor John Odindi (Co-Supervisor) Signed……………….. Date…06/06/2022…

Declaration 1: Plagiarism

I Odebiri Omosalewa Olamide declare that:

1. The research reported in this dissertation is my original work unless otherwise indicated.
2. This dissertation has not been submitted for the attainment of a degree or examination purposes at another university.
3. This dissertation does not contain any data, graphics, and other information from other persons unless duly acknowledged.
4. This dissertation does not contain other persons' writings unless duly acknowledged as such. In cases where written sources have been cited;
    a. Their words have been paraphrased and general information attributed to them has been referenced.
    b. Where exact words have been used, they were placed inside quotation marks and referenced.
5. This dissertation does not contain text, graphics, and or tables directly copied and pasted from the internet unless otherwise sources were duly acknowledged within the content of this dissertation.

Signed[redacted] Date: ….06 June 2022……

Declaration 2: List of Publication and Manuscripts

1. **Odebiri, O**., Odindi, J., & Mutanga, O. (2021). Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review. *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102389.

2. **Odebiri, O**., Mutanga, O., Odindi, J., Naicker, R., Masemola, C., & Sibanda, M. (2021). Deep learning approaches in remote sensing of soil organic carbon: a review of utility, challenges, and prospects. *Environmental monitoring and assessment*, *193*(12), 1-18.

3. **Odebiri, O**., Mutanga, O., & Odindi, J. (2022). Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data. *Geoderma*, *411*, 115695.

4. **Odebiri, O**., Mutanga, O., Odindi, J., & Naicker, R. (2022). Modelling soil organic carbon stock distribution across different land-uses in South Africa: A remote sensing and deep learning approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, *188*, 351-362.

5. **Odebiri, O**., Mutanga, O., Odindi, J., & Naicker, R. (2022). Mapping soil organic carbon distribution across South Africa's major biomes using remote sensing-topo-climatic metrics and Concrete Autoencoder-Deep neural networks. *Science of the Total Environment,* Under Review, Manuscript ID: **STOTEN-D-22-10001**

6. **Odebiri, O**., Mutanga, O., Odindi, J., & Naicker, R. (2022). Evaluation of projected soil organic carbon stocks under future climate and land cover changes in South Africa using a deep learning approach. *Journal of Environmental Management,* Under Review, Manuscript ID: **JEMA-S-22-05916**

Dedication

I dedicate this thesis to God almighty, the giver of life and strength, and to my amazing parents and loved ones.

Acknowledgment

In the city of our God and in the mountains of His holiness, the Lord is glorious and greatly to be praised because He is beautiful in all situations. All acclaim and honour go to God, the giver and sustainer of life, for providing the chance and strength to overcome the challenges of pursuing a Doctorate degree. My profound gratitude to the University of KwaZulu-Natal's School of Agricultural, Earth and Environmental Sciences, for providing me with the opportunity to pursue my studies.

My supervisors: Prof. Onisimo Mutanga and Prof. John Odindi have my deepest gratitude and admiration for their moral support, scientific direction, invaluable critical comments, and mentorship in helping me to scientifically reason and write well. To be honest, I can't thank you both enough for your encouragement and belief in me. I consider myself quite fortunate to be under your guidance, and I pray that God richly blesses you both. I'd like to thank the UKZN-funded Big Data for Science and Society Project (BDSS) as well as the NRF Chair in Landuse Planning and Management (Grant no. 84157), for their financial support of this study.

Special thanks to the staff of the Department of Geography at the University of KwaZulu-Natal: Brice Gijsbertsen (a dear friend and big uncle), Prof. Hill, Dr. Romano Lottering, Dr. Shenelle Sewells, Shanita Ramroop and Mr. Devos, for the love and assistance. To friends and colleague in the Department; Rowan Naicker, Kimara (My baby sister), Mthembeni Mngadi, Adeola, Forbes, Samantha, Lwando Royimani, Collins, Trylee, Serge, and Cecelia; thank you for your support. To my family and friends at the Church on Campus, you all made Pietermaritzburg a home for me. My special gratitude goes to Pastor Shaun and Shanita Ramroop; thanks for your love, hospitality and prayers.

My heartfelt gratitude goes to my family: Mr.& Mrs. Odebri, Bros Dayo and Pastor Odebiri; thank you for your constant prayers and support. To Abimbola John, Sithokozile Precious, Mandy Chibambo, Ifeoluwa Olatunbosun, Bro Bayo, Gedion Tolufashe, Tola Agboola; thanks for your moral support and always being there when I call. To my big brothers: Kunle Olaniyan and Kingsley Echendu; thank you for your support and encouragement during my study. Not forgetting my brother and friend Iyanu Ajileye, thanks for who you are to me. To the Ajileye and Olaniyan family, thanks for making me part of the family; I have never felt otherwise. Indeed, thanks for the love and prayers. God bless you all.

Table of Contents

List of Tables

List of Figures

Acronyms

AEs	Autoencoders
ANN	Artificial neural network
AVHRR	Advanced Very High-Resolution Radiometer
Boruta-DNN	Boruta-Deep neural networks
BPNN	Backpropagation neural networks
CAE-DNN	Concrete Autoencoder-Deep Neural Networks
CI	Confidence interval
CNN	Convolutional Neural Networks
CO_2	Carbon dioxide
D`RAP	Durban Research Action Partnership
DBN	Deep Belief Networks
DDEMs	Data-driven empirical models
DEM	Digital Elevation Model
DL	Deep learning
DMOSS	Durban Metropolitan Open Space System
DNN	Deep neural network
DSM	Digital soil mapping
DVI	Difference Vegetation Index
ELM	Extreme learning machine
ENVI	Environment for Visualizing Images
ESA	European Space Agency
EVI	Enhanced Vegetation Index
GAN	Generative Adversarial Network
gC/m2	Grams of carbon per square meter
GCM	General Circulation Models
GDP	Gross domestic product
GHG	Greenhouse gases
GNDVI	Green Normalized Difference Vegetation Index
GRU	Gated recurrent unit
GtC yr-1	Gigatons carbon per year
IPCC-GPG	Intergovernmental Panel on Climate Change Good Practice Guidance

IPPC	Intergovernmental Panel on Climate Change
ISRIC	International Soil Reference and Information Centre
LCCC	Lin’s concordance correlation coefficient
LIDAR	Light Detection and Ranging
LSTM	Long short-term memory networks
LUCAS	Land Use/Land Cover Area Frame Survey
MFR	Multiple-factor regression
ML	Machine learning
MLP	Multilayer perceptron
MLR	Multiple linear regression
MODIS	Moderate Resolution Imaging Spectroradiometer
MSAVI	Modified Soil Adjusted Vegetation Index
MSE	Mean squared error
NDVI	Normalized Difference Vegetation Index
NIR	Near-infrared
NN	Neural network
NPP	Net Primary Productivity
OCK	Ordinary co-kriging
OK	Ordinary kriging
OLCI	Ocean and Land Colour Instrument
OOB	Out-of-bag
OSAVI	Optimized Soil Adjusted Vegetation Index
PCA	Principal component analysis
Pg	Petagram
PLSR	Partial least square regression
R2	Coefficient of determination
RBF	Radial basis function
RDVI	Renormalized Difference Vegetation Index
ReLU	Rectified linear unit
RF	Random forest
RMSE	Root mean square error
RNN	Recurrent Neural Networks
RS	Remote sensing
RVI	Ratio Vegetation Index

SAEES	School of Agricultural, Earth and Environmental Sciences
SANLC	South Africa National Land-Cover Map
SAVI	Soil Adjusted Vegetation Index
SD	Standard deviation
SFR	Single-factor regression
SFTS	Space-for-time substitution
SGB	Stochastic gradient boosting
SHAP	SHApely Additive exPlanations
SNAP	Sentinel Application Platform
SOC	Soil organic carbon
SOM	Soil organic matter
SRTM	Shuttle Radar Topography Mission
SSPs	Shared Socio-economic Pathways
SVM	Support vector machine
t/h	Tonnes per hectare
TNN	Traditional neural networks
TVI	Transformed Vegetation Index
TWI	Topographic wetness index
UAVs	Unmanned Aerial Vehicles
UNEP	United Nations Environment Programme
VIs	Vegetation indices
WHO	World Health Organization

Chapter One:
General Introduction

1.1. Introduction

Carbon emissions from a range of sources remain a critical influence to the increasing global warming (Laurin *et al.,* 2014). Rising levels of carbon dioxide (CO_2) and other greenhouse gases in the atmosphere are mostly due to anthropogenic activities that include burning of fossil fuels, deforestation, and urbanization (Wolf *et al.*, 2011). Houghton *et al.,* (1996) for instance, predicted a CO_2 emission increase from 7.4 Gigatons (Gt) carbon per year (GtC yr-1) (1Gt = 1 Petagram (Pg) = 1015 g) in 1997 to approximately 26 GtC yr-1 by 2100; while the Intergovernmental Panel on Climate Change (IPPC, 2016) noted that long-term global warming has substantial, widespread, and irreversible repercussions for people and ecosystems. Generally, the dangers of global warming are disproportionately higher for the poor, particularly in resource-constrained areas like Africa (IPPC, 2016). According to World Health Organization (WHO, 2015), climate change is predicted to cause extreme weather events leading to among others drought and crop failures, floods and inundation, as well as prevalence of diseases and pests. WHO (2015), further notes that warmer climates facilitate infectious disease incidences, which will likely result in 250,000 fatalities annually between 2030 and 2050. In addition, IPCC (2021), noted that climate change has exacerbated global economic disparity, and this trend is expected to continue, with the most severe consequences expected in Sub-Saharan Africa where most of the population is reliant on natural and agricultural resources. Consequently, urgent climate change mitigation and regulation strategies are necessary.

Soil organic carbon (SOC) sequestration is one of the most appealing carbon assimilation approaches for mitigating climate change (Padarian *et al.*, 2021). Small changes in SOC could hugely impact global carbon cycle because it is the biggest terrestrial carbon reservoir, accounting for 50 to 80 percent of total world carbon stock (Sahoo *et al.*, 2019; Broderick *et al.,* 2015; Dovey, 2014). The global SOC stock has been estimated to be over 1500 Pg carbon in the top 100 cm of soil, which is approximately double the amount of carbon in the atmosphere and three times the amount stored in terrestrial vegetation (Batjes, 1996; Smith, 2004; Liu *et al.,* 2011). SOC is also an important indicator of soil fertility and land degradation. According to the United Nations Convention to Combat Desertification (UNCCD, 2019), SOC stocks are one of three Land Degradation Neutrality (LDN) indicators that must be continuously monitored and reported on a regular basis. Furthermore, research has demonstrated that SOC influences biological, chemical, and physical processes such as soil water capacity, soil structure, soil cation exchange capacity, and organic matter, making it one of the most important components of vegetation quality and health (Wang *et al.,* 2018; Mondal *et al.,* 2017; Shukla *et al.,* 2006). As such, developing fast, reliable, and accurate procedures for regular quantification, monitoring, and assessment of SOC stocks at regional and global scale is becoming increasingly important (IPCC 2021).

Soil organic carbon stocks and distribution is driven by multiple environmental factors (Levine *et al.,* 2000; Liu *et al.,* 2006; Yoo *et al.,* 2006; Zhang *et al.,* 2008; Afshar *et al.,* 2010). As noted by Bhandari and Bam, (2013), the amount and distribution of SOC within any landscape is determined by multiple variables such as vegetation, temperature, rainfall, topography, land use and management practices. Furthermore, soil organic carbon is generally highest within the first few meters and decreases substantially with increase in depth (Afshar *et al.,* 2010). As indicated by Jobbágy and Jackson (2000), the amount and variability of SOC within the first three-meter depth in the soil is based on the influence of vegetation rather than climate and topography. However, beyond the first three-meter depth, SOC has a stronger relationship with climate and topography rather than vegetation (Jobbágy and Jackson, 2000). Other studies have reported that regardless of the depth, SOC accumulation and distribution is region specific due to peculiar environmental conditions that dictate respective SOC presence (Yigini & Panagos 2016; Sanderman *et al.,* 2017; Minasny *et al.,* 2017; Zhao *et al.,* 2021). These arguments suggest that in-depth understanding of the various environmental variables affecting the formation, distribution and dynamics of SOC is still necessary. Such knowledge is also critical to developing suitable management strategies to improve carbon assimilation as well as assessing the responses and feedbacks of terrestrial ecosystems to climate change (Chaplot *et al.,* 2010). However, obtaining reliable SOC estimates, especially over a vast area, remains a major challenge due to factors such as few sample points, simulation data quality, and the technique used. Hence, cutting-edge solutions that can link environmental variables to SOC processes over large spatial scales are required (Fissore *et al.,* 2017; Fiener *et al.,* 2015; Liu *et al.,* 2011; Chaplot *et al.,* 2010).

Although field-based and laboratory traditional approaches for SOC determination are highly accurate, they are costly, tedious, destructive, time consuming, and often difficult to execute over large areas (Bhunia *et al.,* 2017; Mzinyane *et al.,* 2015). In contrast, remote sensing (RS) approaches offer relatively cost-effective means of quantifying soil properties including SOC. In recent years, RS has become a rich and valuable source of information as it plays a critical role in landscape analysis (Li *et al.,* 2018). According to Hamida *et al.,* (2018), RS approaches are arguably the best strategies for cutting edge and in-depth understanding of the earth systems and other numerous purposes that include long-term climate studies, population evolution analysis, environmental behaviours, and intelligent counteraction of catastrophes. The value of RS can be attributed to among others recent sensor innovations characterized by finer spatial and spectral resolutions, multiplication of image datasets and advances in software/hardware capabilities (Odindi *et al.,* 2016; Mutanga *et al.,* 2015). The use of RS data in form of spectral vegetation indices and bands for instance, permits effective mapping of SOC with good and acceptable accuracies when integrated with geostatistical and machine learning (ML) models. This affords the opportunity to provide updated, consistent, and spatially explicit assessment of SOC and its dynamics, particularly in large and remote areas (Odindi *et al.,* 2016). Additionally, other environmental factors such as topographic and climatic features can be easily derived from Digital Elevation Model (DEM) and WorldClim data respectively. Consequently, the application of remote sensing approaches in quantifying different soil properties is increasingly becoming popular (Gonzalez *et al.,* 2010; Koch, 2010; Zhang *et al.,* 2014).

Since it is proved that SOC has strong correlations with vegetation density, terrain and climatic variables, numerous geospatial models have been explored in relating field-measured SOC to RS derived variables (Wan *et al.,* 2019; Zhang *et al.,* 2019; Richardson *et al.,* 2017; Bhunia *et al.,* 2017; Kumar *et al.,* 2016). Sensor's spectral bands and vegetation indices (VIs), defined with various combinations of visible, near-infrared (NIR) and shortwave reflectance as well as climatic and terrain variables, are the most widely used for retrieving biophysical and biochemical parameters such as SOC (Zhang *et al.,* 2019). Parametric models including multiple linear regression (MLR), ordinary kriging (OK), partial least square regression (PLSR) and principal component analysis (PCA) are commonly used to develop relationships between SOC and RS predictors. However, since the early 2000s, non-parametric models, particularly conventional ML algorithms such as support vector machine (SVM), random forest (RF) and stochastic gradient boosting (SGB) have become prevalent due to their ability to reveal non-linear patterns as well as address issues associated with data dimensionality in fitting models with a large number of predictors (Ma *et al.,* 2019; Zhang *et al.,* 2019). While these models have provided acceptable accuracies, they have also shown to be limited in their ability to extract more complex non-linear abstract elements required for improved predictive models (Wang *et al.,* 2021; Wadoux *et al.,* 2019). Besides, SOC within and between regions exhibit large variability due to their complex blend of organic and inorganic components, hence classical ML models may be unable to accurately predict the behaviour of such a complex phenomenon, especially at large spatial scales (Ma *et al.*, 2019; Padarian *et al.,* 2019; Kumar *et al.*, 2016).

Since 2014, deep learning (DL) models have attracted attention within the RS community due to their ability to automatically extract invariant and abstract features with better discrimination capabilities than geostatistical and traditional ML models (e.g. Singh and Kasana, 2019; Chen *et al.,* 2019; Zhang *et al.,* 2019; Wadoux *et al.,* 2019). Deep learning are algorithms based on neural networks which comprise neurons, also known as units, with many layers that transform input data (e.g. remotely sensed data) to outputs such as categories while learning progressively higher-level features (Schmidhuber, 2015; Litjens *et al.,* 2017). The layers between the input and output are often referred to as "hidden" layers. A neural network containing multiple hidden layers is typically considered as a "deep" neural network—hence, the term "deep learning" (Litjens *et al.,* 2017). Deep learning can be categorized into two major groups namely; supervised and unsupervised (Romero *et al.,* 2015; Li *et al.,* 2018). Supervised learning entails learning with the aid of an algorithm from a well-labelled training dataset and consists of an input and output variable, while unsupervised learning involves modelling the underlying or hidden structure of data and only consist of an input variable (Wittek, 2014).

Although several studies have used DL approaches to analyse RS data (Xu *et al.,* 2019; Zhang *et al.,* 2019), there are two major areas yet to be fully explored: firstly, regression analysis: most DL applications have been primarily used for classification purposes such as LULC classification, object detection, and scene classification (Zhang *et al.,* 2019). This assertion is further supported in the recent DL applications global review by Ma *et al.,* (2019). They noted that, of the 171 journal article published (DL/RS data related) globally between 2008 and 2018, most studies focused on image classification while limited studies focused on the use of DL architectures for regression purposes even though the few studies conducted indicated a considerable level of success. For instance, Zhang *et al.,* (2019), conducted a deep learning-based regression analysis for quantifying above-ground biomass and obtained high accuracy of 94%. Secondly, the use of multispectral sensors, as most DL models are frequently used to analyse hyperspectral and very high-resolution images. Majority of the image data cited in more than 100 recent studies had a spatial resolution finer than 2 m (Ma *et al.,* 2019). This suggests that remote-sensing data with very high spatial resolution benefits more from DL, probably because such data contains rich spatial feature information. Nonetheless, exploring different DL architectures on multispectral data (e.g. Sentinel and Landsat series) could be beneficial to understanding earth systems and processes, as well as permit mapping at an affordable cost especially in resource constraint zones like Africa. Moreover, despite the demonstrated benefits of DL over other standard ML models, research on the implementation of DL-based RS techniques for SOC modelling is scarce, and most extant studies are localized with minimal global impact (Padarian *et al.*, 2020). Taking advantage of the abundance of RS data and the analytical prowess of DL architectures could therefore provide significant potential for rapid, continuous, and reliable national scale SOC estimates. This could be useful in among others informing credible national climate policies and soil management and achieving national total annual and global carbon accounting objectives, as well as IPCC and Kyoto protocol objectives (IPCC 2016; Ndalowa 2014). Consequently, the findings of this study will establish a framework for mapping and monitoring the state of SOC stock distribution across different South African landscapes, as well as how the use of remote sensing especially multispectral data and DL can be integrated into SOC management practices.

1.2. Aim and objectives

This study aimed to explore DL-based approaches for large scale analysis of remote sensing data to predict SOC stocks distribution across South Africa. The objectives of the study were;

1. To provide an overview on the basic and deep learning models in remote sensing of soil organic carbon estimation

2. To review the utility, challenges, and prospects of deep learning approaches in remote sensing of soil organic carbon

3. To map soil organic carbon stocks distribution at a national scale using Sentinel-3 data and a deep learning approach

4. To determine soil organic carbon stocks distribution within South Africa’s major land uses, using a remote sensing and deep learning approach

5. To investigate soil organic carbon distribution within South Africa’s major biomes using remote sensing-topo-climatic metrics and Concrete Autoencoder-Deep Neural Networks

6. To evaluate and project soil organic carbon distribution under future climate and land cover changes in South Africa, using multi-source data and a deep learning approach

1.3. Research questions

1. How accurately can remotely sense data be used to accurately predict large scale soil organic carbon stocks?
2. Do deep learning algorithms perform better than other conventional machine learning techniques in the prediction of large scale soil organic carbon stocks?
3. Which derived spectral vegetation indices and spectral bands best describe SOC stock variability?
4. Which environmental (topo-climate) variables best explain soil organic carbon stocks variability in the study area?
5. How does the selected variable influence the amount of soil organic carbon stocks variability?
6. How accurately can future soil organic carbon distribution be projected using deep leaning-based remote sensing technique?

1.4. Study site description

This study was conducted in South Africa (Figure 1.1), covering an area of 1,221,037 square kilometres. South Africa is the largest country in Southern Africa and Africa's ninth largest country. It is located in a temperate climate zone delimited by the Indian and Atlantic Oceans with Namibia to the northwest, Mozambique and Swaziland to the northeast and east, and Botswana and Zimbabwe to the north and complete enclosure of Lesotho as bounding countries (Department of Environmental Affairs, 2017). The country comprises of nine provinces including Northern Cape, Western Cape, Eastern Cape, North West, Free State, KwaZulu-Natal, Gauteng, Mpumalanga and Limpopo. With an average of 8-10 hours of sunshine per day, South Africa is one of the sunniest countries in Africa. The average daily temperature in summer and winter is about 20°C and 15°C (Scott and Lesch, 1997). Most of the country's interior is relatively a flat plateau with an altitude between 1,000 and 2,100 meters (Mzinyane *et al.*, 2015). Its average annual rainfall is about 464 mm, which is considered low compared to the 786 mm global average (Schulze and Scuttle, 2020). The country's semi-desert, arid and semi-arid conditions means its soil carbon content is lower, compared to countries with abundant vegetation cover (Schulze and Schutte, 2020). The country's major soils morphology is locally classified as lithic, cumulic and oxidic; while the major land uses are varied forms of agriculture that occupy about 79% (Venter *et al.*, 2021).

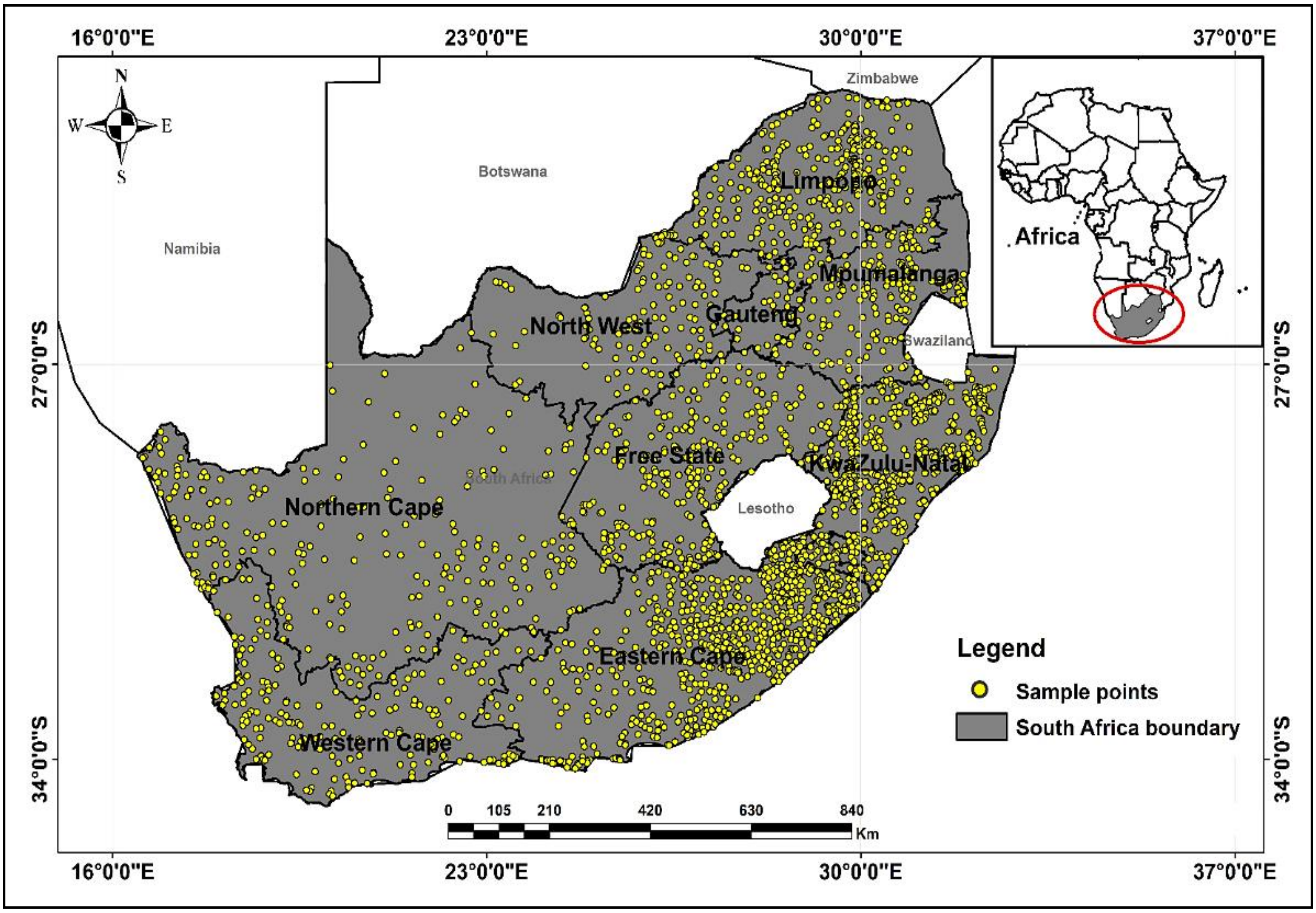


Figure 1.1. The location of South Africa showing the boundaries of the nine provinces and the spatial distribution of soil samples

1.5. Thesis structure

Excluding the general introduction and the synthesis chapters (chapters 1 and 8 respectively), this thesis comprises six research papers that answer each of the research objectives outlined in section 1.2. The literature review and methodologies are entrenched within the mentioned papers. Out of the six research papers, four has been published in peer-reviewed journals while the remaining two are under reviews. Kindly note that each article is presented as a separate chapter within this thesis and is structured in the traditional peer-reviewed article format. Each chapter begins with an introduction and concludes with a link to the next chapter. As a result, there are some commonalities and theory repetition between chapters. This was unavoidable due to the parallel transition of doctrines that serve as the foundation of current scientific knowledge. In this regard, each chapter should be viewed as a self-contained, stand-alone piece of work, but this should not detract from the thesis's overall context.

Chapter Two: This chapter provides a summary on basic and deep learning models used in remote sensing of soil organic carbon quantification. It emphasizes the shift from simple or shallow neural networks (basic) with a single hidden layer to complicated designs with numerous hidden layers currently known as deep learning. Within the context of remote sensing of soil organic carbon, the chapter also outlines the differences, similarities, and limitations of basic networks versus deep learning.

Chapter Three: This chapter provides a comprehensive review of the utility challenges, and prospects of deep learning-based remote sensing approaches for mapping soil organic carbon stocks. In this paper, the history and application of deep learning frameworks in the remote sensing field is discussed, starting with traditional neural networks and other conventional machine learning models. Graphical illustration and in-depth quantitative review with results of the usage of traditional and deep learning models are explained; as it relates to soil organic carbon mapping. Finally, limitations and recommendation together with the future of deep learning based remote sensing of soil organic carbon are examined.

Chapter Four: This chapter focuses on the use of a deep neural network (DNN) to map the distribution of national soil organic carbon stocks using spectral data and vegetation indices obtained from the most recent sentinel series (Sentinel 3) data. The chapter demonstrates the superiority of deep learning models by comparing the results with other traditional neural network and conventional machine learning that include artificial neural network (ANN), random forest (RF) and support vector machine (SVM). Using the SHAP technique, this chapter also provides a solution to the interpretability constraint frequently associated with deep learning frameworks, when compared to other models like RF. This process establishes important DNN model's explanatory variables including sensitive Sentinel 3 band regions and vegetation indices critical for mapping soil organic carbon in the study area.

Chapter Five: This chapter investigates soil organic carbon stocks distribution across major South Africa land uses, using Deep Neural Networks (DNN) and Sentinel-3 satellite data. The chapter highlights how soil organic carbon is significantly influenced by anthropogenic land use change, with intensive and extensive disturbances resulting in considerable soil organic carbon loss. Seven major land uses including grassland, natural forest, cropland, commercial forest, barrenland, shrubland and urban vegetation are evaluated for soil organic carbon stocks distribution, as well as their sequestration rate which is vital for integrated national land-use planning and climate change mitigation.

Chapter Six: This chapter employs remotely sensed topo-climatic data and a hybrid deep learning approach known as Concrete Autoencoder-Deep Neural Networks to map soil organic carbon distribution across major South African biomes (CAE-DNN). The hybrid model's first phase (CAE) performed dimensionality reduction to eliminate redundant input variables and enhance accuracy, while the second phase (DNN) mapped soil organic carbon stocks using the variables chosen. The hybrid model's results were compared to Boruta (Boruta-DNN), a popular classical machine learning feature selection technique, and a standalone DNN model without a variable selection strategy. Also discussed in the chapter is the variability of soil organic carbon stocks and their sequestration rates among nine biomes including Savanna, Grassland, Nama karoo, fynbos, Succulent karoo, Albany thicket, Indian ocean coastal belt, Deserts and Forests. Finally, the chapter provides a guideline to facilitate sustainable SOC stock management within South Africa’s major biomes and indeed other parts of the world.

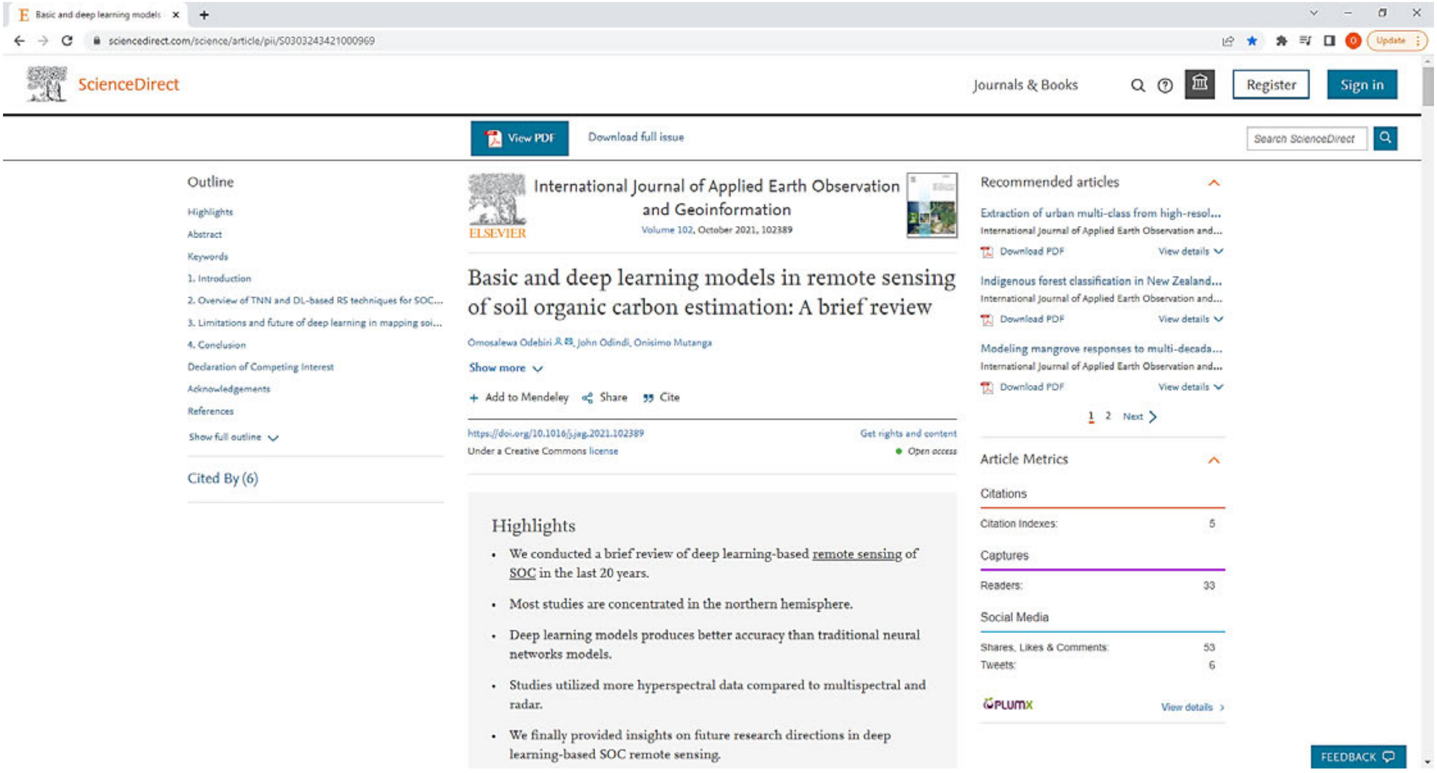
Chapter Seven: This chapter is based on the evaluation of projected soil organic carbon stocks under future climate and land cover changes in South Africa using a deep learning approach. Here, Digital soil mapping (DSM) strategy together with space-for-time substitution (SFTS) processes are used to map and analyse SOC changes induced by projected climate and land cover changes over South Africa between 2021 and 2050. The chapter also evaluates the potential SOC changes between 2021 and 2050 over South Africa's key land uses that include grassland, natural forest, commercial forest, cropland, shrubland, barren land and built-up vegetation. The chapter's first phase uses a Deep neural network (DNN) to estimate current SOC content (2021), while the second phase uses an average of five WorldClim General Circulation Models to project SOC to the future (2050) under four Shared Socio-economic Pathways (SSPs): SSP126, SSP245, SSP370, and SSP585, which represent low, medium, medium to high, and high emission pathways, respectively. Finally, the chapter provides a framework for government and policymakers to evaluate the efficacy of current soil organic carbon management systems.

Chapter Two:

Basic and deep learning models in remote sensing of soil organic carbon estimation: a brief review

This chapter is based on;

Odebiri, O., Odindi, J., & Mutanga, O. (2021). Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review. *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102389.



International Journal of Applied Earth Observation and Geoinformation
Volume 102, October 2021, 102389

Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review

Omosalewa Odebiri, John Odindi, Onisimo Mutanga

https://doi.org/10.1016/j.jag.2021.102389
Under a Creative Commons license

Highlights

- We conducted a brief review of deep learning-based remote sensing of SOC in the last 20 years.
- Most studies are concentrated in the northern hemisphere.
- Deep learning models produces better accuracy than traditional neural networks models.
- Studies utilized more hyperspectral data compared to multispectral and radar.
- We finally provided insights on future research directions in deep learning-based SOC remote sensing.

Recommended articles

Extraction of urban multi-class from high-resol...
International Journal of Applied Earth Observation and...

Indigenous forest classification in New Zealand...
International Journal of Applied Earth Observation and...

Modeling mangrove responses to multi-decada...
International Journal of Applied Earth Observation and...

Article Metrics

Citations

Citation Indexes:	5

Captures

Readers:	33

Social Media

Shares, Likes & Comments:	53
Tweets:	6

Abstract:

Understanding soil organic carbon (SOC) is critical to, among others, atmospherics and terrestrial carbon balance and climate change mitigation. This has necessitated the development of novel analytical approaches to determine SOC. The use of neural network (NN) models for the analysis of landscape biophysical and biochemical properties based on remotely sensed data has become popular in the last decade. This has been facilitated by the proliferation of “big data” from earth observation systems as well as by advances in machine learning (ML). Specifically, the use of traditional neural networks (TNN) and transition to deep learning (DL) frameworks offers considerable improvement in the performance and accuracy of SOC retrieval from remotely sensed data. This chapter seeks to provide a summative assessment of the use of TNN and DL-based remote sensing strategies in SOC estimation, with focus on the progression, application, potential and limitations. The review concludes by providing major challenges impeding the wide adoption of DL frameworks in remote sensing applications of SOC, as well as examining potential directions for future research.

Keywords: Deep Learning, Remote Sensing, Soil Organic Carbon, Traditional Neural Network

2.1. Introduction

Soil Organic Carbon (SOC) is the largest terrestrial carbon reservoir (between 50-80 % of the total terrestrial carbon) and an important means of mitigating climate change (IPCC 2016). It determines a landscape's carbon's source/sink ability and influences soil's physical, chemical, and biological properties (Odebiri *et al.*, 2020a). As such, SOC has attracted significant attention in agriculture, ecology, climate change and sustainable development studies (Wang *et al*., 2018).

Recently, the adoption of remote sensing (RS) based multivariate models in SOC studies has increased significantly (Odebiri *et al.*, 2020b). However, due to the diverse nature of RS data that include high dimensionality, spatio-temporal data volume, and multiple contiguous bands, RS data analysis faces numerous logistical and practical challenges. Additionally, the pre-processing and analysis procedures are often profoundly reliant on the type of model adopted. This has necessitated continuous search for novel analytic strategies to improve the use and performance of RS data and analysis. Central to these advances is the transition from the traditional machine learning (ML) to deep learning (DL) techniques. Unlike physical models that mainly rely on prior knowledge of parameters, DL exploits feature representations learned exclusively from data (Zhu *et al*., 2017). The DL models can enhance learning procedures from complex non-linear relationships between properties, and have demonstrated superiority over geostatistical and other existing ML algorithms (Zhang *et al*., 2019). In existing literature, the term 'DL' and 'ML', have been mistakenly used interchangeably. Whereas this could be justifiable, as DL is a branch of the broader ML, in this review, models that are solely built from neural networks (NN) will be referred as DL and not ML. Also, traditional neural networks (TNN) that have been used in literature for SOC modelling, such as extreme learning machine (ELM), multilayer perceptron (MLP), backpropagation neural networks (BPNN), and radial basis function (RBF) will be included to fully understand the use of DL-based remote sensing techniques in SOC estimation.

Recently, Yuan *et al*. (2020) conducted a broader review on DL applications in among others, crop yield, land covers and vegetation parameters. Generally, most of the existing reviews have focused on the development and application of DL models, computational requirements and the technicalities involved in DL architecture. However, the use of DL models for retrieving environmental parameters, such as SOC have been largely ignored. This is important considering the large body of existing literature that have utilized neural networks (both traditional and deep networks) in SOC modelling (Kuang *et al.*, 2015). As such, DL-based remote sensing of SOC deserves more attention to fully understand the complexity and dynamics of SOC.

2.2. Overview of TNN and DL-based RS techniques for SOC retrieval

Deep learning (DL) are models based on neural networks, otherwise called units with numerous "hidden" layers that transform input data to outputs, while progressively learning higher-level features (Schmidhuber, 2015). Traditional neural network (TNN) differ from DL as they contain shallow layers (i.e. one or two hidden layers) while a neural network containing multiple hidden layers is typically considered as a "deep" - hence, the term "deep learning" (Litjens *et al*., 2017).

A number of studies have successfully incorporated the use of neural network RS-based models for the prediction of soil organic carbon/matter (SOC/M). Whereas the use of neural networks (specifically TNN) models have been largely successful, they are generally unpopular, partly due to high computational demands (Ma *et al*., 2019). This has prompted a shift to other computationally efficient geostatistical and machine learning models such as partial least square regression (PLSR), principal component analysis (PCA), support vector machine (SVM), and random forest (RF) (Mountrakis *et al*., 2011). Around 2014, there was a renewed interest in the utilization of NN structures within the remote sensing community (Ma *et al*., 2019). This is attributed to the availability of richer databases and the emergence of novel technical developments and computational tools such as DeepLearningKit, Microsoft Cognitive Toolkit, Tensorflow and Keras (Zhang *et al*., 2016). Subsequently, TNN models characterized by shallow layers have been developed to encompass multiple hidden layers capable of inducing more representative data features (Litjens *et al*., 2017). Furthermore, the availability of rich repositories has enabled consolidation of various state-of-the-art deep learning frameworks that include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Autoencoders (AEs), Generative Adversarial Network (GAN) and Deep Belief Networks (DBN) (Zhu *et al*., 2017).

2.2.1. Basic learning structures

As aforementioned, several NN structures have been developed to effectively address different types of RS data challenges. Examples of TNN models that have been utilized for SOC/M retrieval include BPNN, RBF, MLP, and ELM (Fidencio *et al*., 2002; Daniel *et al*., 2003). On the other hand, only two (e.g. Tsakiridis *et al*., 2020 and Singh & Kasana, 2019) of the mainstream deep learning (DL) models (i.e. CNN and RNN) have been utilized for SOC/M mapping. Hence, it is necessary to briefly discuss the nature of these frameworks.

The Backpropagation neural network (BPNN), a classical NNs, is characterized by a single hidden layer between the input and output layer and can also contain several nodes or neurons at each layer. It is categorized into forward and backward propagation models, and works by building and initializing the network structure, repeatedly feeding the training data and making predictions with the trained network (Mutanga & Skidmore 2004). BPNN is one of the most popular TNN models that have been used for RS-SOC mapping (Zhao *et al*., 2020). For example, Jaber *et al*., (2011) demonstrated the potential of BPNN for SOC estimation using Hyperion data (400-2500 nm) over a wetland. Recently, Chen *et al*., (2018), proposed a backpropagation neural deep learning (BPN-DL) and compared the results with the traditional BPNN, PLSR, and PCA models. However, Yuan *et al*., (2020) notes that use of BPNN frameworks suffer from sensitivity to network weights and slow convergence toward a state of minimum error.

The multilayer perceptron (MLP) framework is a universal approximator class of feed-forward artificial neural network (ANN) (Falahatkar *et al*., 2016). It was introduced to improve the performance of BPNN structures and a commonly used NN in remote sensing applications (Gupta *et al*., 2016). MLP typically contains input, output, and an intermediate hidden layers. Generally, the input and output layer (X, Y) receive the signal (i.e. data), to make a prediction, while the hidden layer proceeds to store the model parameters (i.e. weight and bias) (Gruszczyński, 2019). MLP is quite flexible as it permits for definition of the number of hidden layers and the neurons within them. A dropout rate function can also be implemented to reduce overfitting and improve accuracy (Xu *et al*., 2019). Kuang *et al.* (2015) constructed an MLP model for SOC retrieval within the visible and near-infrared (305–2200 nm) sections of the electromagnetic spectrum while Chen *et al*., (2019) extended the basic MLP framework to incorporate four hidden layers for SOC prediction. However, the use of MLP framework is constrained by difficulty during training.

The radial basis function (RBF) is a feed-forward NN with outstanding approximation function capabilities, hence popular in among others pattern recognition and function approximation (Broomhead and Lowe 1988). It typically uses the Gaussian kernel function made of three layers; an input layer with neurons that feed the feature variable into the network, a hidden layer of RBF neurons, and an output layer (Fidencio *et al.*, 2002). In SOC/M prediction, Li *et* *al.* (2013a) proposed a soil organic matter (SOM) RBF model using a dataset derived from Advanced Very High-Resolution Radiometer (AVHRR) within the 580-1000 nm, while Li *et al.* (2013b) augmented the RBF framework to incorporate Ordinary Kriging (OK) for SOC retrieval using a spectrometer (350-2500 nm) and a MODIS dataset (400-2300 nm). Compared to BPNN and MLP architectures, the optimal parametrization in RBF is guaranteed due to less complex structure and training procedure (Gautam *et al.*, 2011). However, use of RBF is impeded by lack of proper rules to determine the number of hidden nodes (Yu *et al.*, 2011) and need for higher memory, caused by different NN training system (Samek and Dostal, 2009).

Extreme learning machine (ELM) was proposed for a single hidden layer feed-forward neural network to tackle the issue of slow convergence by other TNN frameworks (Huang *et al.,* 2004). The ELM has a similar structure to other TNN, but with a quicker learning rate (Song *et al.*, 2017). Moreover, ELM has the ability to randomly select the parameters of hidden vectors and can therefore logically calculate the output weights. Hence, the training process and iterations are extremely fast and efficient (Pang and Yang, 2016). The ELM NN has been utilized for different soil property retrieval tasks that include soil temperature, soil moisture, heavy metals, and SOC/M (Lin *et al.,* 2014; Song *et al.* 2017; Yang *et al.* 2019). Despite the fact that ELM trains faster than other TNN models, it is limited by slow evaluation and validation of the trained model (Sirsat *et al.*, 2018).

2.2.2. Deep learning structures

Although the TNN models have been successful for SOC/M retrieval, they are less robust due to their generally shallow structures (Yuan *et al.*, 2020). Furthermore, most of the models are commonly slow to converge during training, and very sensitive to weights, which may affect their convergence (Liu *et al.*, 2018). Consequently, deeper and flexible frameworks, such as RNN and CNN have been developed to address these challenges.

Convolutional Neural Networks (CNN), like other DL algorithms, consists several layers. Within the input and output layers are three major hierarchical structures namely; convolutional, pooling, and fully connected layers (Veres *et al.*, 2015). The convolutional layer is usually placed at the start of the network (i.e. input image) and several local filters can be applied to perform the convolution operation (Veres *et al.*, 2015). The pooling layer, through functions such as max/average-pooling can help reduce high data dimensionality (Wadoux *et al.*, 2019). A simple example of a CNN model output for a given input image (**X**), can be presented as:

$X' = p(ø(W * X+b))$,

Where W = matrix of weights

b = vectors of neuron bias

* = convolutional operator over dimensions' width(w) and height(h) of X

p(.) = pooling function that selects the maximum value of a given input image X, and

ø = activation function usually a rectified linear unit (RELU)

The final convolutional layer returns an X image which can be converted into a vector (i.e. flatten operation), and subsequently given as an input to a fully connected layer which outputs the final result (Wadoux, 2019).

CNN was initially designed to process data in multi-array forms, making it well suited for multi-band RS images (Ma *et al*., 2019). In soil spectroscopy, it was pioneered by Veres *et al*. (2015) who constructed a 2D-CNN model for different soil properties using the Land Use/Land Cover Area Frame Survey (LUCAS) dataset. Later, Padarian *et al*., (2019) transformed the spectral data from the LUCAS database (400-2500 nm) to a 2-D spectrogram to formulate a 2D-CNN model to multi-task and predict SOC and other soil properties simultaneously. Generally, CNN frameworks are more efficient than other neural networks and can automatically detect important features in a data without human intervention (Dotto *et al*., 2018). As such, CNN is regarded as the most powerful and popular mainstream DL model in RS applications (Yuan *et al*., 2020). However, its efficacy is limited by among others, computational cost and the need for a large training dataset (Somarathna *et al.,* 2017).

Recurrent Neural Networks (RNN) is a widely used robust supervised learning model primarily used for sequential problems (Rodriguez *et al*., 1999). Unlike standard feed-forward TNN, RNN contains a loop structure that allows exhibition of dynamic temporal behaviour for sequential data processing (Ma *et al*., 2019). Generally, RNN comprises of three major parts with several hidden layers. The nodes in the input sequence are progressively added into the RNN to derive the corresponding output sequence (Yuan *et al*., 2020). Moreover, given the inbuilt RNN memory, it can easily recall critical information beneficial for a reliable future output prediction (Singh & Kasana, 2019). As such, RNN often performs well with sequential input tasks such as time-series applications. For instance, a sequential input data such as a hyperspectral image of one pixel can be represented as $H = < i_1, i_2, i_3>$; while the hidden layer of the RNN can be expressed with the following equation;

$$h_t = \begin{cases} 0, & if\ x \geq 1 \\ \omega(h_{t-1}, i_t), & othewise \end{cases}$$

where h_t = the current hidden state,
h_{t-1} = previous hidden state,
i_t = current input
ω = activation function which can be hyperbolic tangent or sigmoid function. The final output layer (H) of the RNN can then be a sequence or a single output (Singh & Kasana, 2019).

RNN models are commonly limited by inability to learn and store information for long, due to the deep feed-forward networks they generate. To resolve this problem, specialized memory units, i.e. long short-term memory networks (LSTM) and gated recurrent unit (GRU) have been developed to augment the networks (Ma *et al*., 2019). Although RNN, and by extension LSTM and GRU have been adopted in RS, only Singh & Kasana (2019) has adopted it in SOC retrieval. Other mainstream DL frameworks such as Deep Belief Networks (DBN), Autoencoders (AEs), and Generative Adversarial Network (GAN) have been used for different RS applications.

2.3. Limitations and future of deep learning in mapping soil carbon

Although DL models have generally been successful in RS based applications, including SOC mapping, a number of challenges still hinder its effective usage, these include; (1) the large number of samples requirement that are often difficult to acquire (e.g collection of soil samples or laboratory derivation of their carbon content equivalent is tedious and labour intensive), (2) computational time, (3) interpretability, (4) end-user technical knowhow, (5) requirement for large storage capacity, (6) constrain of RS dataset acquisition by cloud cover, (7) limited ground data, resulting to missing satellite data, and (8) tendency for model over-fitting. In addition, there has been a lack of result-accuracy consistency which could be due to differences in calibration datasets or variation in study sites; as SOC is dynamic in nature and varies from one area to another (Somarathna *et al*., 2017; Dotto *et al*., 2018; Odebiri *et al*., 2020a).

Despite the limitations and challenges highlighted above, it is important to note that the future of DL-based remote sensing technique for SOC modelling is promising. For instance, the use of multispectral sensors with DL models for SOC modelling is an area that is yet to be fully explored. A review of literature shows that majority of the existing DL-based remote sensing studies for SOC analysis mostly used hyperspectral sensors as opposed to multispectral and radar. Several authors have argued that the low spatial-spectral resolution in commonly used

RS imagery is the major impediment to the adoption of multispectral sensors (Ma *et al*., 2020). Nevertheless, the launch of the new generation multispectral sensors such as Worldview-2 and 3(31 cm), Rapid-eye, Sentinel-2 and 3, with improved spatial-spectral resolution and strategically positioned bands sensitive to soil could be beneficial to DL SOC retrieval task. Furthermore, future DL SOC studies can adopt image fusion to improve accuracy (Liu *et al*., 2018). Fusion techniques combines two or more different datasets to obtain a single higher spatial-spectral resolution image and hence improve accuracy. For instance, fusing a low resolution multispectral image with a high resolution panchromatic image known as pan-sharpening; or the fusion of hyperspectral-multispectral, hyperspectral-radar, and multispectral-radar (Huang *et al*., 2015). Additionally, the use of Unmanned Aerial Vehicles (UAVs) remains largely unexplored in SOC mapping. UAV platforms are relatively cheaper and flexible to use as they permit the end-user an opportunity to manually select the optimum weather conditions for image acquisition (Guo *et al*., 2020; Angelopoulou *et al*., 2019). Moreover, other mainstream DL models such as Autoencoders (AEs) and Deep Belief Networks (DBN) that have proven effective in other retrieval task are yet to be fully tested for SOC mapping. They could be beneficial to SOC mapping, particularly in assessing the performance of existing DL models like the CNN and RNN. The issue of limited sample size could also be addressed by the transfer learning technique proposed by Goodfellow *et al*., (2016). Transfer learning generally works by adjusting the parameters of a DL model previously trained on a large dataset with smaller samples for optimal performance in a new task (see Yuan *et al*., 2020 for further details on transfer learning)

2.4. Conclusion

This study provides a brief review on the use of NN and transition to DL for SOC analysis using remotely sensed data. Specifically, the study highlights the characteristics of major TNN and DL models, their use in SOC analysis as well as their strengths and limitations. Whereas this review is by no means exhaustive, it notes the major limitations of the TNN as among others; difficulty during training, need for higher memory and slow evaluation of the trained model. On the other hand, DL is limited by among others the need for large training samples, computational cost, technical knowhow, input and output data volume and overfitting. It is anticipated that the emergence of both commercial and freely available remotely sensed "big data" will lead to wider adoption of novel approaches like DL in SOC analysis. Finally, a comprehensive review is also required to unpack the history and development of DL based-remote sensing of SOC.

2.5. Summary

The essential principles, differences, similarities, and limits of TNN and DL-based remote sensing of SOC were discussed in this chapter. It stresses the transition from simple or shallow neural networks with a single hidden layer to complex architectures with multiple hidden layers referred to as deep learning. Lastly, the study provided a summative assessment of the predominate drawbacks and prospects surrounding the use of TNN's and DL within the remote sensing of SOC. Given the anticipated wider adoption of DL frameworks within SOC analyses, the next chapter actively set out to provide an in-depth quantitative review that contextualizes the use of deep learning-based remote sensing approaches within the field of soil organic carbon mapping. The study systematically tracks the theoretical and geographical evolution as well as the challenges and opportunities associated with the topic.

Chapter Three:

Deep learning approaches in Remote Sensing of Soil Organic Carbon; a review of utility, challenges, and prospects

This chapter is based on;

Odebiri, O., Mutanga, O., Odindi, J., Naicker, R., Masemola, C., & Sibanda, M. (2021). Deep learning approaches in remote sensing of soil organic carbon: a review of utility, challenges, and prospects. *Environmental monitoring and assessment*, *193*(12), 1-18.

Visit Nature news for the latest coverage and read Springer Nature's statement on the Ukraine conflict

Published: 15 November 2021

Deep learning approaches in remote sensing of soil organic carbon: a review of utility, challenges, and prospects

Omosalewa Odebiri, Onisimo Mutanga, John Odindi, Rowan Naicker, Cecilia Masemola & Mbulisi Sibanda

Environmental Monitoring and Assessment **193**, Article number: 802 (2021) | Cite this article

552 Accesses | Metrics

Abstract

The use of neural network (NN) models for remote sensing (RS) retrieval of landscape biophysical and biochemical properties has become popular in the last decade. Recently, the emergence of "big data" that can be generated from remotely sensed data and innovative machine learning (ML) approaches have provided a platform for novel analytical approaches. Specifically, the advent of deep learning (DL) frameworks developed from traditional neural networks (TNN) offer unprecedented opportunities to improve the accuracy of SOC retrievals from remotely sensed imagery. This review highlights the use of TNN models and their evolution into DL architectures in remote sensing of SOC estimation. The review also highlights the application of DL, with a specific focus on its development and adoption in remote sensing of SOC mapping. The review concludes by highlighting future opportunities

Abstract:

The use of neural networks (NN) models for remote sensing (RS) retrieval of landscape biophysical and biochemical properties has become popular in the last decade. Recently, the emergence of “big data” that can be generated from remotely sensed data and innovative machine learning (ML) approaches have provided a platform for novel analytical approaches. Specifically, the advent of deep learning (DL) frameworks developed from traditional neural networks (TNN) offer unprecedented opportunities to improve the accuracy of SOC retrievals from remotely sensed imagery. This review highlights the use of TNN models and their evolution into DL architectures in remote sensing of SOC estimation. The review also highlights the application of DL, with a specific focus on its development and adoption in remote sensing of SOC mapping. Findings from literatures show that majority of published articles are concentrated in the Northern Hemisphere, i.e., China (38), Iran (11) and USA (9), while Africa as a continent had only 4 publications. Results also reveal that most studies adopted hyperspectral data particularly spectrometers compared to multispectral data. The TNN (90%) models were mostly used in literature compared to DL (10%); however, DL models produced better median accuracy (~93 %) compared to the TNN (~85%). The review concludes by highlighting future opportunities for the use of DL frameworks for the retrieval of SOC from remotely sensed data.

Keywords: Deep Learning, Remote Sensing, Hyperspectral, Multispectral, Radar, Soil Organic Carbon, Climate Change

3.1. Introduction

Challenges associated with carbon emission and climate change have become pervasive globally (Yuan *et al*., 2020; Padarian *et al*., 2020). Consequently, to mitigate the adverse effects of the changing climate, monitoring and understanding carbon assimilation pools is paramount (Odindi *et al*., 2015; Mutanga *et al*., 2015). Soil organic carbon (SOC) accounts for the largest proportion of terrestrial carbon (i.e. between 50-80 %); hence a critical pathway for climate mitigation (Zhao *et al*., 2020; IPCC 2016). Soil organic carbon is a dynamic entity, which plays fundamental roles that include facilitating the wellbeing of a functional soil and providing a major carbon source/sink within the global carbon cycle (Lamichhane *et al.*, 2019). Furthermore, SOC influences a soil's biophysical and chemical properties, strengthens its structural characteristics and improves its water and nutrient holding capacity (Angelopoulou *et al.*, 2019). As such, SOC studies have attracted great interest in the fields of, among others, soil science, precision agriculture, natural and commercial forestry, land degradation, climate change and sustainable development (Sahoo *et al*., 2019; Wang *et al*., 2018; Aryal *et al*., 2017).

In recent decades, SOC studies using remote sensing (RS) approaches based on multi-variate models have increased significantly (Madileng *et al.*, 2020; Odebiri *et al*., 2020b; Hamida *et al*., 2018). This is attributed to myriad advantages offered by RS over traditional approaches. According to Mngadi *et al.* (2019) and Khanal *et al*. (2018), RS methods are cheaper, faster, non-destructive, and cover a wide spatial range, making them useful at both a micro and macro spatial scales. According to Xiao *et al.,* (2019), the value of RS datasets can be attributed to the ability of platforms and sensors to fully capture different landscape data, useful in investigating the range, trend, and conversion of SOC dynamics at all levels. However, due to the varied nature of remotely sensed data, that include data dimensionality, spatio-temporal and spectral information, and numerous proximate bands, RS analysis is confronted with a range of analytical challenges. Consequently, the remote-sensing community is continuously committed to advancing innovative analytical approaches to optimize the use and output of RS data (Li *et al.*, 2018; Masemola and Cho 2019).

Central to these advances is the emergence of deep learning (DL) techniques that have proven to be novel analytical tools in many fields. Specifically, the application of DL algorithms for environmental remote sensing has rapidly increased over the last 10 years (Padarian *et al*, 2019a). Unlike other physical models that majorly depend on preceding knowledge of parameters, DL approaches capitalize on the representations of features solely derived from the data (Zhu *et al*., 2017). These approaches are capable of enhancing learning procedures, particularly from complex non-linear correlations between different environmental properties and have demonstrated their superiority over geostatistical and other traditional machine learning (ML) approaches (Zhang *et al*., 2019; Minh *et al*., 2018). Whereas the term 'DL' and 'ML' have commonly been used interchangeably in literature, this review was restricted to neural networks (NN) DL approaches that have been adopted for SOC modelling e.g., the extreme learning machine (ELM), multilayer perceptron (MLP), backpropagation neural networks (BPNN), and radial basis function (RBF).

In recent years, several reviews on DL-based remote sensing techniques have emerged. For example, Ma *et al*., (2019), Liu *et al.*, (2018), Li *et al*., (2018), Zhu *et al*., (2017), and Zhang *et al.*, (2016) performed extensive reviews on the application of DL for image classification, 3D modeling, data fusion, image segmentation/registration and image preprocessing. Other reviews have focused on specific thematic areas. For instance, Shen *et al*. (2018a) and Di Noia and Hasekamp (2018) conducted reviews on hydrology and atmospheric aerosol analysis, respectively. In a recent study, Yuan *et al*. (2020) performed an in-depth review on the use of DL in environmental remote sensing. Their review covered among others, radiation, temperature, ocean colour, evapotranspiration, snow cover, rainfall, particulate matter, aerosol, land cover analysis and vegetation biophysical parameters. To date, existing reviews on DL have commonly dwelt on their development, usage, computational demands, and architecture. However, the use of DL models to retrieve environmental parameters like SOC remains un-reviewed. Furthermore, SOC is a unique, complex, and dynamic environmental parameter due to its presence in various particulate, labile, humic, recalcitrant, and microbial forms in the soil (Odebiri *et al.,* 2020a). Hence, it is necessary to review DL-based remote sensing of SOC to fully determine how its merits can be harnessed to better understand its complexity and dynamics. In this regard, this study sought to conduct a comprehensive review on the applications of DL-based RS strategies for SOC retrieval. Specifically, the review dwells on the progression of traditional neural networks (TNN) to state-of-the-art DL architectures and its use for remote sensing mapping of SOC. The review also highlights on the major sources of RS data, their technical attributes (i.e., spatial, temporal, and spectral resolution) and their use in the retrieval of SOC. It concludes by providing insights on how DL can be effectively adopted for SOC remote sensing.

Deep learning (DL) models are based on hierarchical progressive higher level "hidden" layers of neural networks that convert input data to outputs (Schmidhuber, 2015). The major difference between traditional neural network (TNN) and DL is that TNN contains one or two hidden layers, commonly referred to as “shallow layers” while multiple layers are considered a “deep” neural network—hence, deep learning” (Litjens *et al*., 2017).

Several studies have adopted a range of neural network models to predict soil organic carbon/matter (SOC/M) using remotely sensed data. Literature has predominately focused on traditional neural networks, with very few contemporary studies utilizing deep learning techniques. One of the earliest studies was conducted in Brazil, by Fidencio *et al*., (2002) who developed a Radial basis function (RBF) NN model from NIR-reflectance spectrometer (1000-2500 nm) data to estimate SOC with relatively high accuracy (R^2 = 0.91, RMSE = 0.25%). Thereafter, Daniel *et al*., (2003) constructed an artificial neural network (ANN) model for SOM prediction in Thailand using VIS-NIR spectrometer (400–1100 nm) with an R^2 = 0.86 accuracy. Whereas the adoption of neural network (particularly TNN) models was been largely efficacious, there was paucity in literature on their use, mainly attributed to high computational demands and lack of adequate data (Ma *et al*., 2019). Hence, the RS community shifted to other geo-statistical and machine learning approaches that include the partial least square regression - PLSR, principal component analysis - PCA, support vector machine - SVM, and random forest - RF that require less computational power and generally produce acceptable accuracies (Mountrakis *et al.*, 2011; Hamida *et al*., 2018). However, about five years ago, a renewed focus on the adoption of NN structures in remote sensing applications emerged (Ma *et al*., 2019). This was attributed to technical developments that include a proliferation of richer databases and ground-breaking computational tools like the DeepLearningKit, Microsoft Cognitive Toolkit, Tensorflow and Keras (Zhang *et al*., 2016; Ciresan *et al*., 2012). Hence, TNN models with shallow layers were established to incorporate numerous hidden layers that could induce more representative data features (Litjens *et al*., 2017). Also, the growing number of rich repositories have enabled an assemblage of numerous novel deep learning frameworks that include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Autoencoders (AEs), Generative Adversarial Network (GAN) and Deep Belief Networks (DBN) (Hinton, 2012; Goodfellow *et al*., 2014; Chen *et al*., 2014; Zhu *et al*., 2017; Minh *et al*., 2018; Hao *et al*., 2018).

Numerous NN structures have been established to handle a range of remotely sensed data challenges. Some of the typical examples of traditional neural network (TNN) models that have been utilized to specifically determine SOC/M are BPNN, RBF, MLP, and ELM. On the other hand, only CNN and RNN, considered as mainstream DL models, have been adopted for SOC/M mapping. The section below highlights on their use and basic structures.

The BPNN is among the most widely adopted TNN models in RS-SOC mapping (Zhao *et al*., 2020). Mouazen *et al*., (2010) for instance compared the performance of BPNN, PLSR and PCA algorithms for SOC estimation using a mobile spectrophotometer (350-2500 nm). They found that the BPNN model (R^2 = 0.84) outperformed the PLSR (R^2 = 0.80) and PCA (R^2 = 0.78) models. Jaber *et al*., (2011) used BPNN to map SOC based on Hyperion data at 400-2500 nm on a wetland and obtained an R^2 = 0.79, RMSE = 3.3 t ha^{-1} accuracy. In a recent investigation, Chen *et al*., (2018), using a spectrometer (1000-2500 nm), developed the backpropagation neural deep learning (BPN-DL) by tuning and selecting 22 hidden layers and their corresponding neurons. The approach significantly improved the results; as the BPNN-DL model (R^2 = 0.95) performed better than the BPNN, PLSR and PCA models with R^2 = 0.88, 0.87, 0.85, respectively. However, the BPNN framework has been criticized for sensitivity to network weights and time consuming convergence to a state of minimal error rate (Yuan *et al*., 2020).

The radial basis function (RBF), a feed forward NN, has remarkable approximation function capabilities, hence popular in a range of applications (Broomhead and Lowe 1988). As shown in Fig 1, it uses a three layered (i.e., input, hidden and output) Gaussian kernel function (Fidencio *et al.*, 2002). The RBF has been adopted in many retrieval tasks, including SOC/M prediction. For instance, Li *et al.* (2013a) suggested a SOM - RBF model from Advanced Very High-Resolution Radiometer (AVHRR) within the visible (580) and the near-infrared (1000 nm) range. The RBF model produced the smallest prediction errors when compared to MLR and regression kriging (RK) with an RMSE of 21.80, 28.04 and 28.05 g.kg^{-1}, respectively. Furthermore, Li *et al*. (2013b) and Li *et al*. (2016) incorporated Ordinary Kriging (OK) into the RBF framework to retrieve SOC from spectrometer and MODIS data at 350-2500 nm and 400-2300 nm, respectively. Compared to BPNN and MLP architectures, the optimal parametrization in RBF is guaranteed because its training procedure and structure are less complex (Gautam *et al*., 2011). Nevertheless, RBF is limited by inadequate rules to effectively determine hidden nodes and higher computational demands (Yu *et al*., 2011; Samek and Dostal, 2009).

The multilayer perceptron (MLP) framework is a feed-forward artificial neural network (ANN) known as universal approximators developed to improve BPNN performance and a commonly adopted NN in remote sensing (Falahatkar *et al*., 2016; Gupta *et al*., 2016). MLP is characterized by at least three layers i.e., input, output, and a hidden intermediate layer (Figure 3.1c). MLP has been used in many aspects of RS research, including SOC retrieval. For example, Leone *et al.* (2013) developed an MLP model to predict SOC (R^2 = 0.89) in Italy using a spectrometer (350–2500 nm). In another study, Kuang *et al.* (2015) developed an MLP model to retrieve SOC using a spectrometer within 305–2200 nm. The accuracy of the MLP model (R^2 = 0.90) was superior to the PLSR model (R^2 = 0.81). Similarly, Chen *et al.*, (2019) improved the rudimentary MLP-DL framework to include four hidden layers, producing a better result than the PLSR model for SOC prediction with an R^2 = 0.92 and 0.80 respectively. However, the efficacy of MLP is impeded by strained training.

The extreme learning machine (ELM) was developed to deal with slow convergence that characterize existing TNN frameworks (Huang *et al.*, 2004). The ELM is similar to TNN (Figure 3.1d), but with faster learning rate (Song *et al*., 2017). The ELM NN has been adopted in among others soil temperature and moisture, metallic composition, and SOC/M (Deng *et al.*, 2015; Ding *et al*., 2015; Lin *et al.*, 2014). Song *et al.* (2017) for instance, evaluated the performance of ELM in comparison to other geostatistical models (RK, OK) using Landsat 8 OLI data (433-2300 nm). The ELM model produced the least error rate (RMSE = 1.402 g.kg^{-1}) in comparison to RK (RMSE = 1.974 g.kg^{-1}) and OK (RMSE = 2.071 g.kg^{-1}). Hong *et al*., (2018) calibrated an ELM model for SOM prediction using a spectrometer (350–2500 nm). Superior predictability was observed in the use of ELM model (R^2 = 0.83) relative to support vector machine (SVM) model (R^2 = 0.82). Similarly, Yang *et al.* (2019) utilized an ELM model for SOM retrieval within a paddy soil in the middle-lower Yangtze Plain of China. The ELM model (R^2 = 0.89) outperformed other models such as SVM (R^2 = 0.86), Cubist (R^2 = 0.78), and PLSR (R^2 = 0.76). Although ELM has better training efficiency than existing TNN models, it suffers from slowed appraisal and authentication of the trained model. In most applications, evaluation speed is more important than the training speed (Sirsat *et al*., 2018).

Whereas the above mentioned TNN models have been successfully adopted for determining SOC/M under varied landscapes, they are less robust in their fitting ability, attributable to their generally shallow structures (Yuan *et al*., 2020). Moreover, they are commonly characterized by slower convergence during a training process, and high sensitivity to weights that influence convergence (Huang *et al*., 2004). This has led to the development of deeper and more flexible frameworks, specifically the RNN and CNN.

CNN, similar to other DL algorithms, commonly contains several layers (Goodfellow *et al*., 2016). The CNNs are typically characterized by convolutional, pooling, and fully connected hierarchical layers within the input and output layers (Figure 3.1e) (LeCun *et al*., 2015). CNN has been adopted for soil spectroscopy using a Land Use/Land Cover Area Frame Survey (LUCAS) dataset by Veres *et al.* (2015) with an RMSE = 0.3841 g.kg^{-1} accuracy. Also, Padarian *et al*., (2019c) achieved a R^2 = 0.94 accuracy from the LUCAS dataset; their 2D-CNN could multi-task by predicting different soil properties, including SOC simultaneously. Similarly, Tsakiridis *et al*., (2020) utilized the LUCAS database (400-2500 nm) to develop a novel framework, which used a localized multiple-channel 1-D CNN to successively estimate different soil properties (SOC) with an R^2 = 0.86 accuracy. Generally, the CNN frameworks are superior to other NN, with an ability to automatically determine vital features in a dataset (Dotto *et al*., 2018), hence most popular DL model in RS (Yuan *et al*., 2020). Nevertheless, its shortcomings include higher computational cost and requirement for large training data (Somarathna *et al.,* 2017).

RNN is a popular supervised learning model mainly used for sequential problems (Rodriguez *et al*., 1999). It is comprised of numerous hidden layers within three major parts (Figure 3.1f) and has demonstrated it superiority in performing tasks that involve sequential inputs including time-series applications. However, because of deep feed-forward networks, it is limited by short learning and information storage periods. Hence, long short-term memory networks (LSTM) and gated recurrent unit (GRU) have been established to support the networks (Ma *et al*., 2019). A detailed description of LSTM and GRU can be found in Hochreiter and Schmidhuber, (1997) and Cho *et al*. (2014). Whereas RNN, including LSTM and GRU have been used in RS applications, literature shows that that it has only been applied by Singh & Kasana (2019) in SOC retrieval. Singh & Kasana (2019) developed an RNN-LSTM model to predict different soil properties including SOC in Europe using hyperspectral data generated from a LUCAS dataset (400-2500 nm) and obtained a better result (R^2 = 0.94) compared to other linear and conventional ML models. Although conventional DL frameworks like Deep Belief Networks (DBN), Autoencoders (AEs), and Generative Adversarial Network (GAN) have been adopted in RS, no known study had been conducted for SOC/M retrieval at the time of this review.

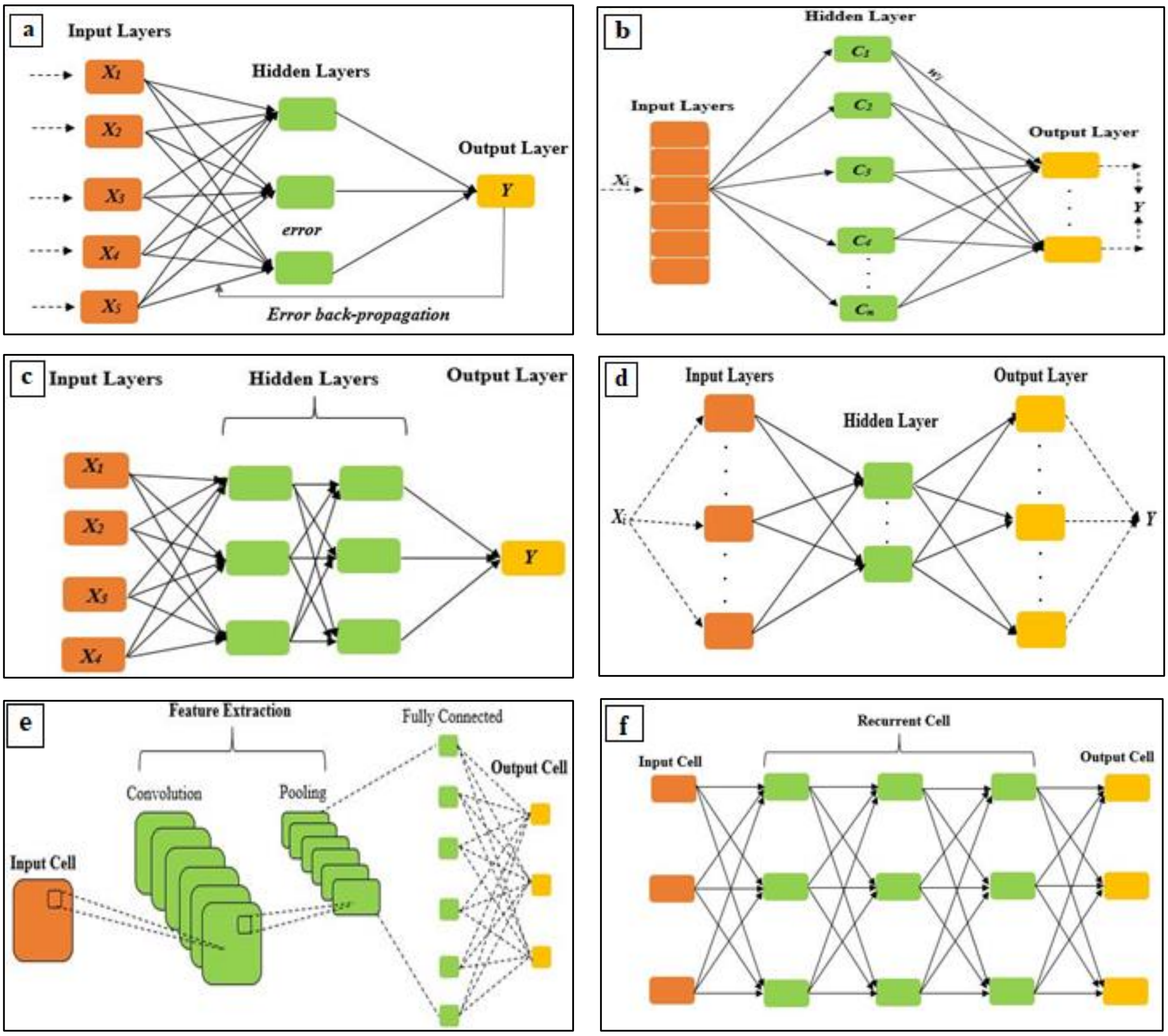


Figure. 3.1. Basic structures of TNN and DL architectures for SOC/M retrieval; (a) BPNN, (b) RBF, (c) MLP, (d) ELM, (e) CNN, and (f) RNN

3.2. Methodology

In this study, we examined the applications of Deep learning-based remote sensing approaches for SOC retrieval in published literature from the year 2002 to 2020. A literature search was conducted using the two mainstream scientific libraries (i.e., Scopus and Web of Science). We then developed some criteria to identify published papers that primarily focused on DL remote sensing techniques for SOC retrieval. These criteria included; (a) the literature must have keywords such as “Deep Learning”, “Neural Network”, “Remote Sensing”, “Soil Organic Carbon” or “Soil Organic Matter” as their essential or auxiliary subject; (b) the title, abstract and keywords of the prospective literature must contain in whole, the predefined keywords as highlighted in the first criterion and (c) the literature must be written in English language and published in a scientific peer-reviewed journal (de Araujo Barbosa *et al*., 2015). On this premise, we utilized the following query keywords; “Deep learning” OR “Neural networks” AND "Remote sensing" AND [“Soil organic carbon” OR “soil organic matter”]. Figure 3.2 presents an illustration of the literature search and the selection process. The initial search resulted in 1731 published papers. Subsequently, book chapters, grey literature, extended abstracts, and technical notes were excluded and duplicates via the endnote database removed (Naicker *et al.*, 2019). Considering that Scopus and Web of Science may be unable to retrieve a significant number of related articles, particularly the year 2020 papers, we performed a manual Google scan for the latest published articles and searched other recently published reviews for related works (Xiao *et al*., 2019; Lamichhane *et a*l., 2019; Angelopoulou *et al*., 2019). At the end of the search process, 95 papers were retrieved for further analysis. To understand the current trend and status of the use of DL models in the retrieval of SOC within the RS field, the 95 papers were sorted and arranged into graphical representations. This included (1) the global distribution of the published articles; (2) the number of peer-reviewed papers published annually; (3) the usage frequency of different NN models and (4) the category (i.e. hyperspectral, multispectral and radar) and platforms (i.e. laboratory spectroscopy, *in situ* spectroscopy, space-borne, airborne and unmanned aerial systems-UAS) of RS data used.

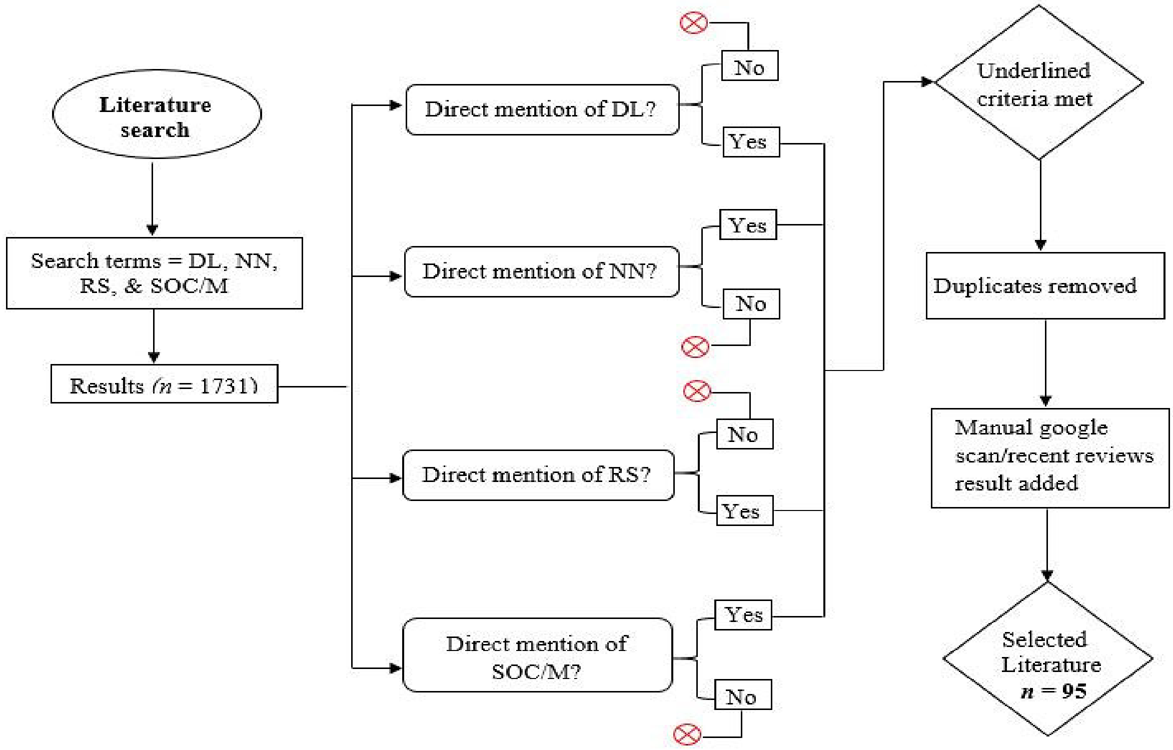


Figure 3.2. Schematic of literature and selection procedure (modified from Naicker *et al.*, 2019; de Araujo Barbosa *et al.*, 2015)

3.3. Results

3.3.1. Number and spatial distribution of publications

Results showed that there has been a steady increase in the direct usage of NN-RS-based techniques for SOC estimation. Figure 3.3 shows the distribution of published articles globally from 2002 to 2020. The majority of published articles are concentrated in the Northern Hemisphere, i.e., China (38), Iran (11) and USA (9). Only four articles have been published in the whole of the Africa continent; largely attributed to the high computational resource requirements for NN processing and the limited resources in acquiring RS data (Yuan *et al* 2020).

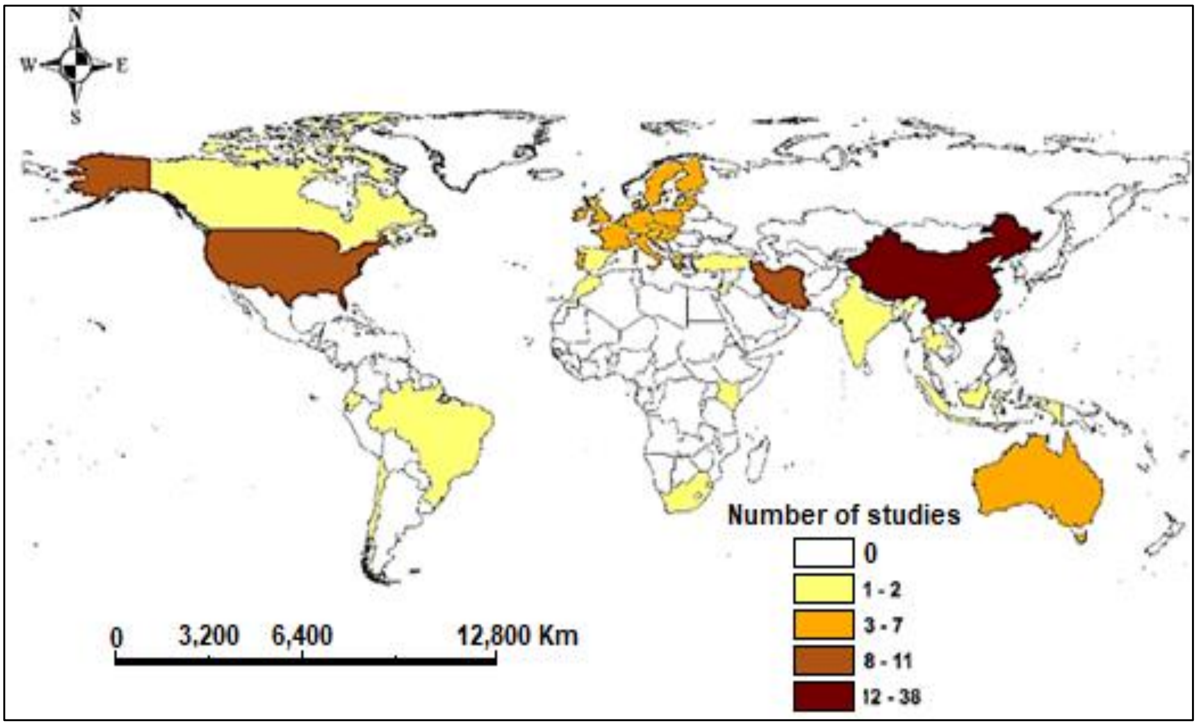


Figure 3.3. Global distribution of the selected published articles on the use of Deep Learning approaches in Soil Organic Carbon mapping

Figure 3.4 presents the number and the trend of publication from 2002 to 2020. Generally, the number of publications has been gradually on the rise. From the year 2002 to 2012 (10 years), there were nine publications using NN models. A significant change in the number of NN model usage was noted in 2013, as eight publications using different NN frameworks were published. The year 2019 showed the highest number (24) of publications while the year (2020) has 10 articles.

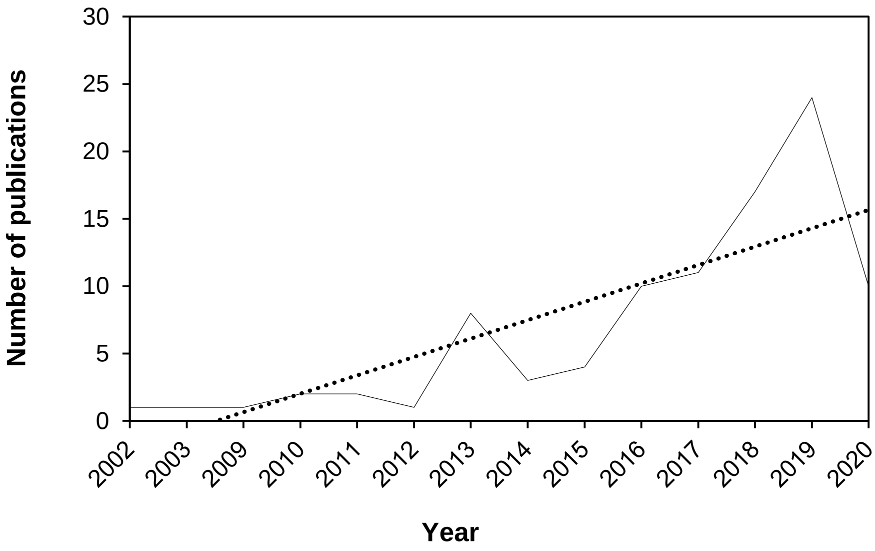


Figure 3.4. Number of published articles from 2002 to 2020 (the dotted line shows a positive trend)

3.3.2. Neural Networks and remotely sensed data used

To understand the progression of deep learning (DL) techniques in the analysis of RS data in relation to SOC prediction, the neural networks (NN) models were divided into traditional neural network (TNN) and DL models. Figure 3.5 shows the use of TNN and DL in mapping SOC using different forms of remotely sensed data. It also shows the sensors, platforms, and the average resolution of remotely sensed data that were used in the selected articles. The TNN models were mostly used across all RS data categories (i.e., Hyperspectral, Multispectral and Radar). A total of 54 hyperspectral studies were conducted with 45 for TNN and 9 for DL. TNN also dominated the multispectral and radar studies with 33 and 32, respectively, while two studies were conducted using DL for both multispectral and radar (Figure 3.5a). The predominant use of TNN in comparison to DL structures can be attributed to lack of interest DL frameworks until 2014; which was the beginning of the application of DL architectures in the RS field (Ma *et al* 2019). Another conceivable explanation could be that numerous scientists were yet to acquaint themselves with the technical know-how and computational skills required to calibrate DL frameworks. The first ever DL-based RS technique for SOC mapping was conducted in 2015 by Veres *et al*. (2015) using the CNN framework. This created the foundation on which other DL related SOC studies have been conducted. In terms of the types of sensors used, Figure 3.5b shows that a large proportion of published work comprised of spectrometers — laboratory/*in situ* — (46 studies), followed by the Shuttle Radar Topography Mission (SRTM) Digital Elevation model (DEM) (34 studies) and the Landsat sensors (25 studies). Sensors such as Hyperion, WorldView and QuikBird, that were used once are classified as others. A closer look at the sensor types shows that new generation freely available multispectral sensors, such as Landsat 8 and Sentinel (1,2 and 3) with moderate resolution, are yet to be adequately tested for the modelling SOC using the DL technique. Figure 3.5c shows the frequency of use of the different RS platforms. Spectrometers (46 studies) were the most popular, followed by space-borne sensors (33 studies) while airborne (4 studies) and unmanned aerial systems (UAS) (1 study) were the least used. As shown in Figure 3.5d, images with high resolution of less than 10m (58 studies), were the most used in all the articles investigated. This shows that DL algorithms perform better with RS data characterized by high spatial-spectral resolution. The rich information contained in these images enable DL to extract more representative features that could benefit SOC prediction (Ma *et al*., 2019).

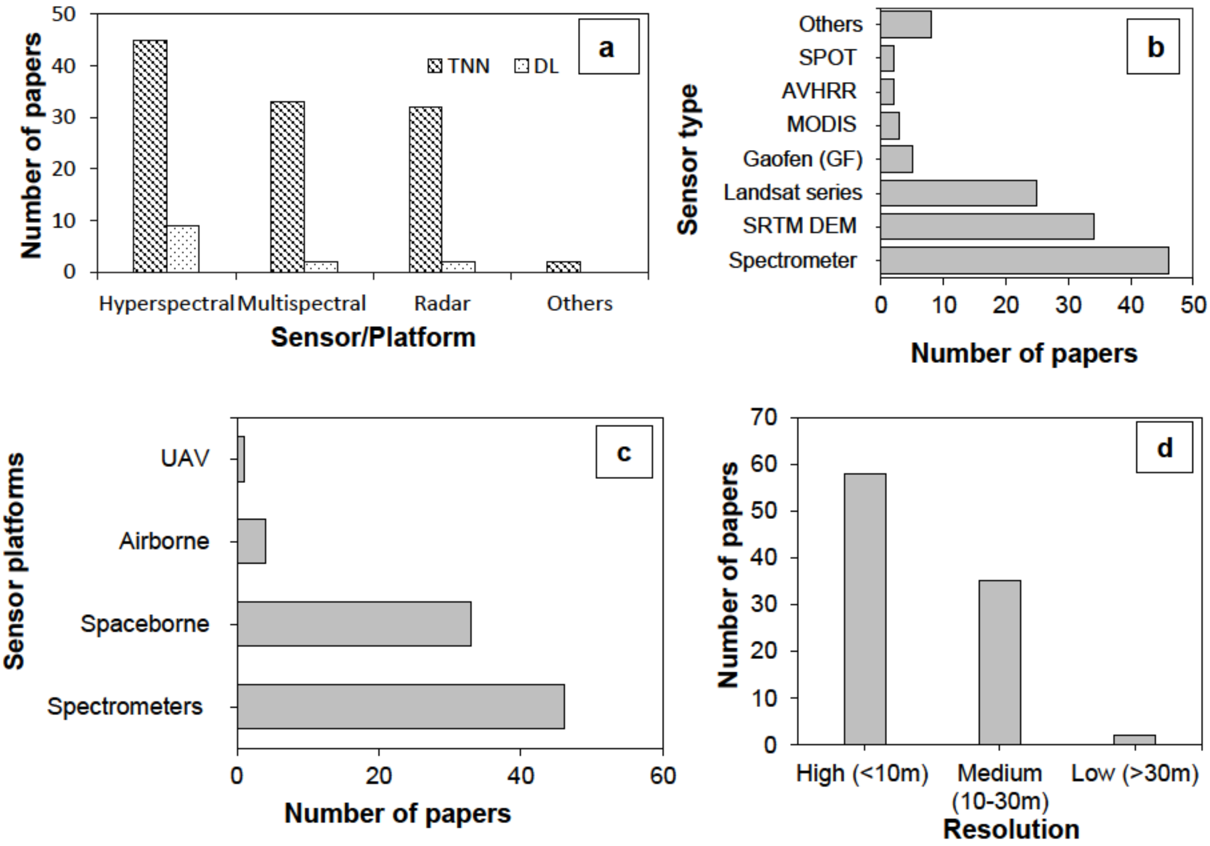


Figure 3.5. (a) The use of traditional neural networks (TNN) and deep learning (DL) in mapping soil organic carbon across different remote sensing data categories, (b) the frequency of use for sensor type; sensors that were used just once are classified as others (c) the frequency of use for sensor platform, (d) Image resolution of reviewed articles grouped into high, medium, and low.

Figure 3.6 shows the different types of TNN and DL used in the selected articles for review. Four TNN models in order of usage were MLP (42), BPNN (34), RBF (10) and ELM (6), while the CNN (9) and RNN (1) were the only DL models used for SOC estimation. Refer to section 3.1. for detailed description of the models.

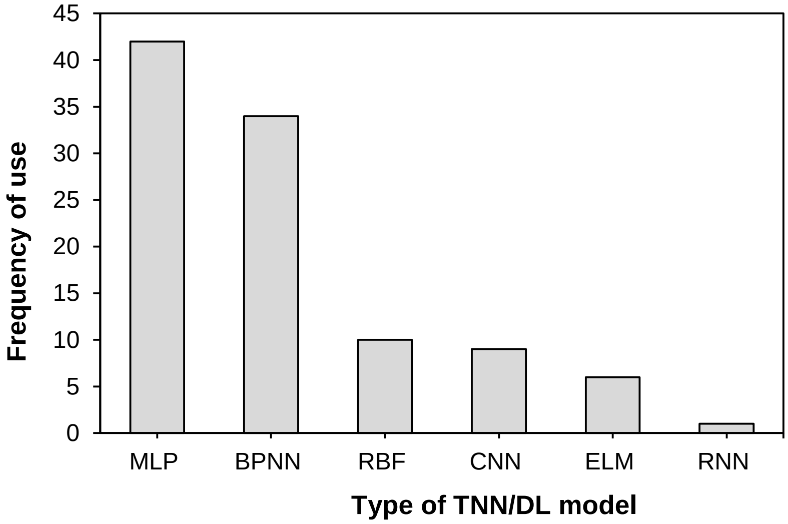


Figure 3.6. The usage of traditional neural networks (TNN) and deep learning (DL) models.

3.3.3. Overall accuracy assessment using the TNN and DL models

Accuracy assessment is an integral part of any modelling technique, as it informs the degree to which a model is reliable in a regression or classification task. Figure 3.7a depicts the overall performance of TNN and DL models within the examined literature. In addition, it shows the accuracy variance across the different RS data types used (i.e., hyperspectral, multispectral and radar). Overall, the DL models with ~93 % median regression accuracy performed better than the TNN models with ~85% median accuracy. The superiority of DL models to TNN models in SOC estimation has been demonstrated by several studies (Yuan *et al.*, 2020). For instance, Xu *et al.*, (2019a), developed a DenseNet and LeNet CNN models for SOC and compared their performance to that of a classical BPNN structure. Using the coefficient of determination (R^2) as a comparison measure, the results indicated that the two CNNs (R^2 = 0.907, R^2 = 0.902) outperformed the BPNN model (R^2 = 0.835). In another study, Taghizadeh-Mehrjardi *et al.* (2020) modified the traditional MPL model (R^2 = 0.75) to incorporate more hidden layers and derived better results for SOC retrieval (R^2 = 0.83). Furthermore, Figure 3.7b shows the accuracy distribution across different RS data types. Regardless of the methods used (i.e., TNN or DL), hyperspectral studies had the highest median regression accuracy (~87 %), followed by multispectral (~80 %), and radar (~76 %). The variance in accuracies can be attributed to the spatial-spectral resolution differences among the RS data types.

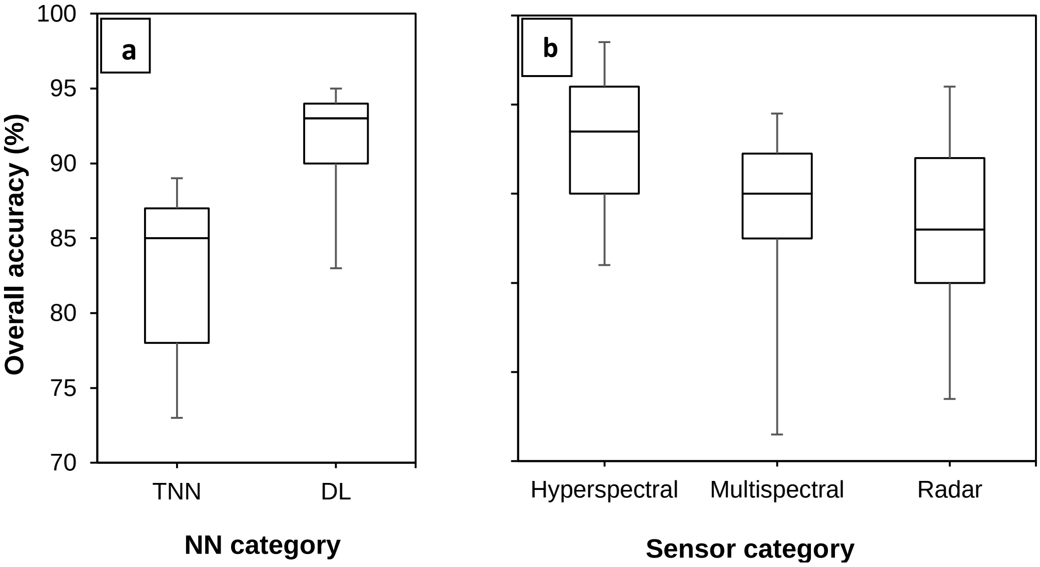


Figure 3.7. (a) Overall accuracies for TNN and DL models; (b) overall accuracies across RS data types (hyperspectral, multispectral, radar)

3.4. General discussion

In this section, we discuss the existing DL-based RS data categories (i.e., Hyperspectral, Multispectral and Radar) and platforms used for SOC estimation in reference to publications over time. We also review the limitations encountered and offer recommendations and future opportunities.

3.4.1. Hyperspectral RS and DL for SOC estimation

Hyperspectral data are generally characterized by numerous spectral bands ranging from 350 nm to 2500 nm (VNIR-SWIR region). They span across different RS platforms from Laboratory-*in situ* spectroscopy to space-borne, airborne, and recently the use of UAS (Angelopoulou *et al*., 2019). From the literature examined, hyperspectral data are the most used RS data type for SOC prediction (Figure 3.5a). This is due to their high spatial-spectral resolution as most of the reviewed hyperspectral studies had an average of 2m spatial resolution. Moreover, DL models have been demonstrated to maximally extract additional associated information from hyperspectral data, which is valuable for different image analysis tasks (Ma *et al*., 2019). Several studies (e.g., Margenot *et al*., 2020; Yang *et al*., 2019; Wijewardane *et al*., 2018) have been conducted using DL-based hyperspectral data for SOC retrieval. These studies began with laboratory and *in situ* spectroscopy using spectrometers (Fidencio *et al*., 2002; Daniel *et al*., 2003). For instance, Janik *et al*., (2009) conducted an *in situ* SOC spectra measurement using an ATR-FTIR Spectrometer in the mid-infrared region (2500-2500 nm) and the resultant BPNN model obtained good (R^2 = 0.93) results. Rossel & Behrens (2010) explored the potential of laboratory spectroscopy (ASD FieldSpec3 spectrometer) to predict SOC in the visible and near-infrared region (350-2500 nm) of the electromagnetic spectrum and the results could explain about 89% of SOC distribution.

The use of space-borne and airborne hyperspectral data has also increased in recent years. They provide the advantage of mapping SOC over larger and inaccessible areas (Odebiri *et al*., 2020). A number of studies have demonstrated the potential of space-borne and airborne hyperspectral sensors using DL models. For example, Jebar *et al*. (2011) predicted SOC using a Hyperion data and a BPNN model with good accuracy (R^2 = 0.789). Dai *et al*. (2014) and Li *et al*. (2016), using MODIS data, successfully developed a RBF framework for the prediction of SOC with a correlation coefficient (R^2) of 0.75 and 0.84, respectively. On the other hand, Zizala *et al*. (2017) used an airborne Spectrographic imager and MLP NN to predict SOC and obtained good (R^2 = 0.88) accuracy. Similarly, with the aid of an airborne AISA Eagle sensor and MLP NN, Zhang *et al*. (2019) estimated SOC stocks in a wetland and obtained an R^2 = 0.90 accuracy. However, to date, there are no known studies that have utilized hyperspectral UAS platform and DL algorithms for SOC/M prediction. The use of UAS system and DL could be beneficial for SOC estimation due to their high spatial resolution and the possibility for scheduled flight plan in accordance with optimal weather conditions.

Overall, the use of DL-hyperspectral data on different platforms for SOC prediction is dependent on the magnitude and cost of study (Meng *et al*., 2020). The use of DL algorithms has proven to perform well with hyperspectral data (Figure 3.5a). DL also produces better accuracy than TNN models and reduces data dimensionality associated with Hyperspectral sensors better than TNN, allowing researchers to investigate SOC in greater detail. However, further research is required to obtain a full understanding of the capabilities and potential of DL. Moreover, hyperspectral data are still expensive, particularly over large areas, and difficult to obtain in many regions, especially Africa (Mutanga *et al*., 2015). Also, laboratory spectroscopy still requires crushing and sieving of soil samples before spectral measurement, which is often tedious and time consuming (Sibanda *et al*., 2015). Nonetheless, hyperspectral data and DL models have been successful in mapping SOC. Additionally, the potential of forthcoming hyperspectral sensors such as PRISMA, HyspIRI and EnMAP with better spatial-spectral and temporal attributes could further benefit the use of DL models in mapping SOC.

3.4.2. Multispectral RS and DL for SOC estimation

Earth observation multispectral space-borne data started with Landsat sensors in 1972 with the launch of Landsat 1. Commonly, multispectral sensors acquire data from high temporal and moderate spatial resolutions using a small number of broad spectral bands across the electromagnetic spectrum (VNIR-SWIR) (Naicker *et al.*, 2019). Multispectral data offer the opportunity for mapping different soil properties at a larger spatial extent than hyperspectral data. In addition, most multispectral data are less expensive while some are freely available, making them ideal for resource constrained zones like Africa (Mngadi *et al*., 2020; Ogbodo *et al*., 2019). Ayoubi *et al*. (2011), Liu *et al*. (2013), Were *et al.* (2015) and Mirzaee *et al*. (2016) led some of the first NN-based SOC estimation studies using spectral band combinations from multispectral data. They generated different vegetation indices from Landsat TM/7 ETM (450-2351 nm) used as predictor variables for SOC and obtained acceptable (R^2 = 0.50—0.70) results. Notwithstanding, majority of the existing studies have noted that low spatial-spectral resolution, together with large swath widths are some of the challenges hindering reliable prediction of different soil properties using DL (Liu *et al*., 2013).

The emergence of new generation multispectral space-borne sensors, with improved spatial and spectral resolution has provided the opportunity for better soil properties retrieval using DL approaches (Zhang *et al*., (2019). Although yet to be fully explored, a number of studies have highlighted the effectiveness of new generation multispectral space-borne in SOC/M retrieval. For example, Zhang *et al*. (2019), conducted a comparative study between a high resolution 63-band AISA hyperspectral data (400-980 nm) and two space-borne multispectral sensors including an 8-band WorldView-2 (400-1040 nm) and a 4-band QuickBird (450-900 nm). They resampled the hyperspectral data to match the specific-spectral attributes of the two space-borne sensors and developed a SOC model using a MLP model. Results obtained indicated not much variance in accuracy (R^2 = 0.90, 0.90, and 0.86 respectively), thus, demonstrating the potential of new generation multispectral space-borne sensors in soil properties mapping. A few airborne and UAS platforms have also been utilized in SOC modelling. For instance, Khanal *et al*. (2018) utilized an airborne multispectral Leica ADS80 digital camera (420-900 nm) together with a 3-layered classical BPNN model to predict SOC and obtained a good (R^2 = 0.64) accuracy. Guo *et al*. (2020) acquired images through an UAS using MicaSense RedEdgeTM 3 camera (475-840 nm). The calibrated model (BPNN) produced (R^2 = 0.80) accuracy.

Regardless of the successes documented above, low resolution, large swath widths, cloud cover, need for geometric and atmospheric corrections as well as low signal-to-noise ratio are some of the challenges facing multispectral data usage (Mirzaee *et al.*, 2016). Furthermore, more DL-based multispectral data investigations are required. This is because majority of the multispectral application reviewed in this study utilized the traditional neural network (TNN). Additionally, many new generation multispectral sensors are yet to be adequately utilized. For instance, freely available sensors like Sentinel 2 (443-2190 nm) and the recently launched Sentinel 3 (400-1020 nm) with improved spatial-spectral resolution and strategically positioned bands including red-edge are yet to be tested.

3.4.3. Radar RS and DL for SOC estimation

Radar data is valuable in generating different topographic variables that are associated with the formation and distribution of SOC stocks (Arogoundade *et al*., 2019). RADAR functions by transmitting a microwave signal towards an object and detects the back-scattered radiation (Minasny *et al*., 2016). Due to radar's long wavelength, it can penetrate canopy cover and thin cloud, as well as generate data in all weather conditions. Furthermore, Radar is highly sensitive to soil conditions, including surface roughness and soil moisture, which benefits SOC estimation (Odebiri *et al*., 2020b).

From the results (Figure 5a), it is evident that Radar is the least adopted RS data type for SOC retrieval tasks using the DL approach. The majority of the radar data (SRTM DEMs) within the reviewed literature were used to complement other RS data types (hyperspectral and multispectral) by generating auxiliary environmental variables such as slope, aspect and elevation in order to improve accuracy (e.g. Wadoux *et al*., 2019a; Hateffard *et al*., 2019; Wu *et al*., 2019; Falahatkar *et al*., 2016). However, few DL-based remote sensing studies have utilized radar data in isolation for SOC prediction. For instance, Bodaghabadi *et al*. (2015) predicted SOC in Iran by deriving 15 different environmental variables from a radar derived DEM (10m) and calibrated a MLP model to obtain a good result (R^2 = 0.88). Similarly, Minasny *et al*. (2016) proposed a cost-effective SOC mapping using the SRTM DEM with a spatial resolution of 30.7 m to generate different terrain covariates. An excellent agreement (R^2 = 0.87) between the observed and predicted SOC was recorded using different multivariate models, including DL (BPNN).

RS data, such as Light Detection and Ranging (LIDAR) and the freely available Synthetic Aperture Radar (SAR) Sentinel-1, are yet to be utilized for DL-based SOC retrieval. Compared to SRTM DEMs, LIDAR offers improved high-spatial resolution data, which are particularly suitable for DL models (Laurin *et al*., 2014). Nevertheless, the high cost of procuring LIDAR, ground data prerequisite for calibration combined with the impracticability in some remote territories are some of the barriers hindering its wide application (Odebiri *et al*., 2020b). Sentinel-1 on the other hand can serve as an alternative to LIDAR in DL RS-based SOC mapping owing to its free availability and its relatively high resolution.

3.4.4. Limitations in DL-based remote sensing techniques for SOC estimation

Although DL models have proven successful in many remote sensing applications including SOC modelling, there are several challenges that hinder their effective use. After a systematic review, we noted that large sample size requirement, computational processing time, interpretability, end-user technical knowhow, large storage capacity requirement and tendency for over-fitting are some of the factors limiting the use of DL architectures. Moreover, there is inconsistency in the results/accuracies where these techniques have been employed. For instance, whereas Ayoubi *et al*. (2011) achieved a good result ($R^2 = 0.84$) using Landsat 7 ETM (450-2351 nm) and a MLP model in Iran, Mirzaee *et al*. (2016) produced a lesser accuracy ($R^2 = 0.63$) using the same data and model in the same country (Iran). This could be due to variation in the selection of calibration data, as well as SOC variation in the study areas as SOC is a dynamic phenomenon that may vary from region to region (Odebiri *et al*., 2020a). As such, there is currently no universally accepted calibration method for SOC retrieval (Lamichhane *et al*., 2019).

Subsequently, many authors (Tsakiridis *et.al* 2020; Wadoux, 2019b; Dotto *et al*., 2018; Somarathna *et al*., 2017) have argued that the use of DL architectures requires a large sample size (>100) to produce acceptable accuracies. Hence, the bigger the sample size, the better the model accuracy. Jordan and Mitchell (2015) also argued that the size of the dataset used for training a DL model is essential, given the large numbers of parameters to fit. Although there is no clear standing rule as to the ideal sample size for training DL models, many studies have established a significant connection between the data sample size and accuracies obtained. For instance, Odebiri *et al*. (2020a) compared the accuracy of ANN and five other models using a small dataset (81 soil samples). Results showed that the ANN ($R^2 = 0.768$) was outperformed by other machine learning models such as Random Forest ($R^2 = 0.84$), Stochastic Gradient Boosting ($R^2 = 0.802$ and Support Vector Machine ($R^2 = 0.79$), due to their ability to handle small sample size data. Similarly, Padarian *et al*. (2019b) demonstrated the importance of sample size to DL, by training a CNN model on two different sample sizes (20,000 and 390 soil samples). Although the CNN model with the higher sample size produced better results than Cubist and PLSR (R^2 = 0.88, 0.79 and 0.35 respectively), it yielded the same result as Cubist (R^2 = 0.79, 0.79) when used on a smaller sample size (390 samples). Conversely, in a study conducted by Chi *et al*., (2018), BPNN model outperformed other models such as multiple-factor regression (MFR), partial least square regression (PLSR) and single-factor regression (SFR). The BPNN model produced the lowest RMSE (2.78 g/kg, 3.67 g/kg, 3.85 g/kg, and 3.88 g/kg respectively), despite using a relatively small samples size (91 soil samples). As a result, it is worth noting that the use of big sample size may not necessarily be the only factor that influences the result. Other factors such as the proficiency of use on the part of the end-users, especially in the tuning of hyper-parameters could also be critical to the success of DL models.

Considering the significance of model interpretability, which alludes to how one can easily comprehend the procedure a model uses to arrive at a result, the DL models are quite complex and less transparent (Were *et al*., 2015). Unlike simpler methods, for example linear models, with well-defined interpretability, DL models are difficult to interpret due to their inability to reveal the functional relationships between spectral information and soil properties (Padarian *et al*., 2020). This could hamper proper understanding of the underlying predictors in the model development (Rossel and Behrens, 2010). Furthermore, computational time and cost is another shortcoming associated with the usage of DL frameworks (Wadoux, 2019b). Wijewardane *et al*. (2016) argued that the best practical model for any retrieval or classification task will be the one that produces a relatively high accuracy and requires less computational cost and time. In this regard, Xu *et al*. (2020) examined the computational time it took for several multivariate models to predict SOC and found that the average time for the neural networks (NN) model was much higher compared to other simpler methods**.**

3.4.5. Recommendation and future opportunities

DL models require a large sample size due to their profound and complex nature. However, most studies are often limited by restricted field samples (Somarathna *et al*., 2017). Furthermore, the number of samples in RS datasets are often constrained by cloud cover and inadequate ground data, resulting in missing satellite data (Yuan *et al*., 2020). This challenge could be resolved by applying the Transfer Learning technique suggested by Goodfellow *et al*., (2016). Transfer learning for environmental RS can be region-based or data-based (Masemola *et al*., 2020). This approach works by adjusting DL model parameters of a formally trained large dataset with smaller samples for optimum implementation on the new task (see Yuan *et al*., 2020 for an extended explanation on transfer learning). In addition, the application of DL mainstream models is yet to be fully tested in SOC mapping. From the results of this review, it is evident that only CNN and RNN have been used. Models such as Autoencoders (AEs) and Deep Belief Networks (DBN) are yet to be tested, despite proving efficient in other retrieval tasks. For instance, Shen *et al*. (2018b) successfully predicted large scale ground surface particulate matter from a MODIS data using DBN and obtained an $R^2 = 0.87$ accuracy. Shao *et al*. (2017), with the aid of AE model, predicted forest above-ground biomass using a synergy of LIDAR, Sentinel-1, and Landsat 8 OLI (433-2300 nm) data and obtained an $R^2 = 0.81$ accuracy. As such, future studies can consider the use of these models in SOC modelling.

A closer look at the literature reviewed in this study showed that a large proportion of the studies utilized hyperspectral than multispectral and radar data (Figure 3.5a). Majority of the literature has cited lower spectral resolution as the main reason hindering the use of multispectral sensors (Ma *et al*., 2020; Naicker *et al*., 2019). The recent launch of commercial multispectral sensors like the Worldview-3 (31 cm) and Sentinel-2 and 3, characterized by better spectral resolution and strategically positioned bands could be beneficial to DL models in SOC retrieval. Worldview-3 is a high spatial resolution multispectral sensor (1.24m), with less than a day temporal resolution (Naicker *et al*., 2019). It possesses a panchromatic and short-wave infrared resolution of 31-cm and 3.7m respectively, hence can be useful in a wide range of DL-based RS applications, including SOC mapping (Kruse *et al*., 2015). For instance, Hively *et al.* (2018) mapped crop residue and tillage intensity over a farmland in Eastern Shore of Chesapeake Bay, USA, based on Worldview-3 image data and obtained an $R^2 = 0.94$ accuracy. However, the high cost of obtaining these images has hindered their frequency of use, especially in resource constrained regions (Wang *et al*., 2016). Conversely, Sentinel-2 and 3 are freely available and cover virtually all potential areas of interest with relatively high spatial-spectral resolution (Mngadi *et al*., 2019). Consequently, more DL-based multispectral investigations need to be conducted as these sensors could offer a cost-effective option to DL-based RS methods for SOC mapping.

Future studies can incorporate the fusion of different RS data types to improve accuracy. Image fusion in RS applications is usually aimed at obtaining a single image that simultaneously possess a high spectral-spatial resolution (Huang *et al*., 2015). The resultant image from a fusion technique can be derived from similar, or different RS data types, for example, combining a lower and higher spatial resolution multispectral and panchromatic images, respectively i.e. pan-sharpening and combining hyperspectral/multispectral, hyperspectral/radar or multispectral-radar (Huang *et al*., 2015). Although a number of fusion based RS application have been conducted for different soil properties mapping (e.g., Gao *et al*., 2017; Lin *et al*., 2020), these studies have been conducted using other traditional ML and linear algorithms. From our review, out of the 95 studies investigated, only one study (Xu *et al*., 2019b) attempted the fusion technique for SOC mapping using a MLP model by combining two hyperspectral images (i.e., attenuated total reflectance Fourier-transform mid-infrared spectroscopy (FTIR-ATR) (2500–25000 nm) and laser-induced breakdown spectroscopy (LIBS) (200-1000 nm). The result ($R^2 = 0.83$) of the fusion data was better than the FTIR-ATR ($R^2 = 0.48$) and LIBS ($R^2 = 0.58$) individually. Attributable to the possibility of generating high spatial-spectral resolution images, we believe subsequent studies using the fusion technique could help in closing the gap of the cost required to acquire high-resolution images favourable to DL models. Additionally, the use of UAS platforms for SOC mapping remains unexplored. The UAS platforms are cheaper and offer an opportunity for use at optimal data acquisition conditions (Guo *et al*., 2020; Angelopoulou *et al*., 2019, Odebiri *et al.*, 2021).

3.5. Conclusion

A systematic survey on the application of TNN and the mainstream DL-based RS techniques for SOC quantification is presented in this study. The number of publications over time and the different types of RS data and platforms used were summarized. Investigations revealed that literature using remote sensing and deep learning techniques in estimating soil organic carbon has largely increased in recent years. However, further research, specifically targeting the effectiveness of different sensors and platforms in estimating SOC is required. In addition, this study provides insight into ways of improving the use of DL architectures. The use of transfer learning, which could address the issue of small sample size, use of improved multispectral sensors (Sentinel-1, 2 and 3, LIDAR, Wordview-3) and versatile platforms (such as UAS), together with the fusion of different RS data types are some of the research avenues future studies should explore.

3.6. Summary

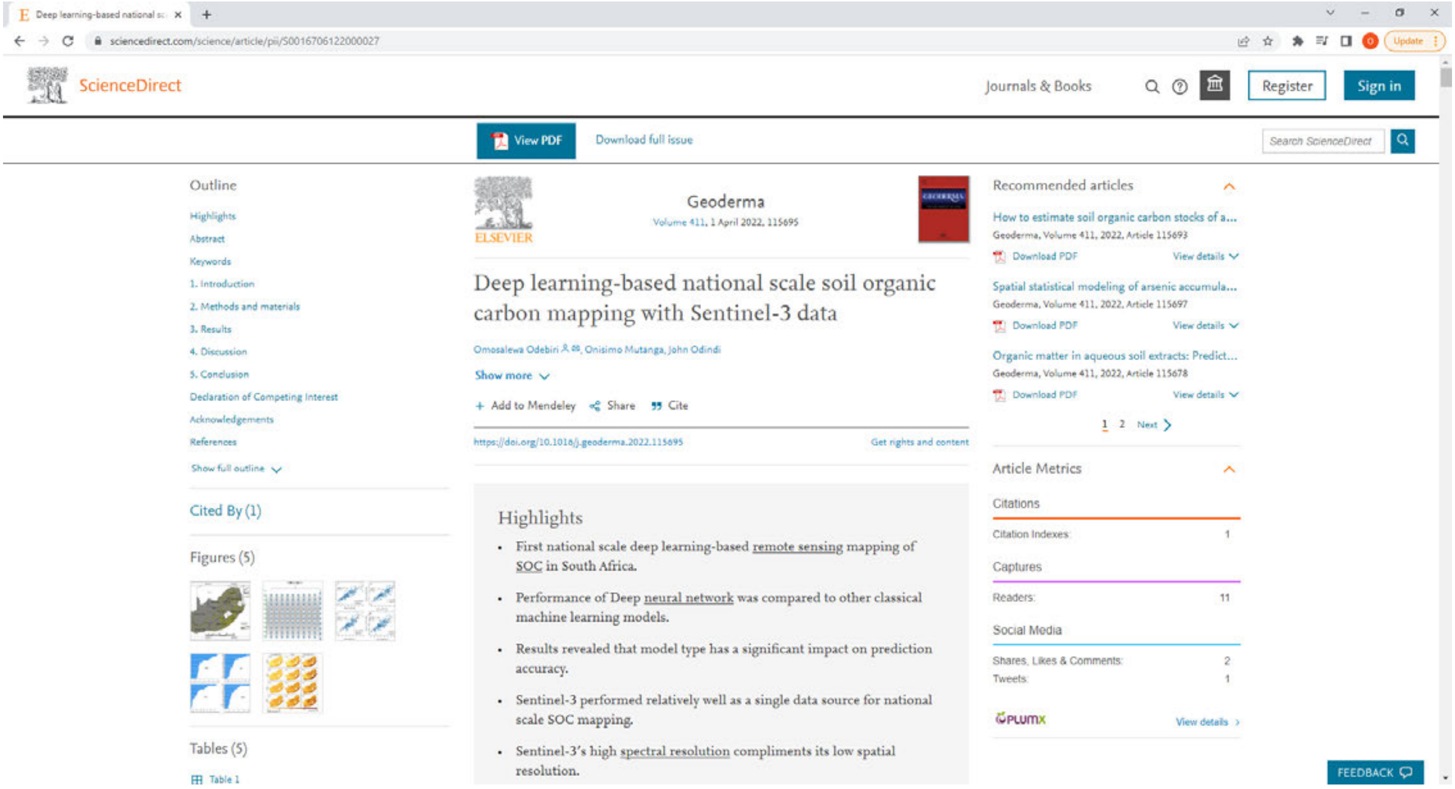
This chapter provided a thorough examination of the use of TNN and other major DL-based RS approaches for SOC quantification. Investigations found that the majority of prior DL-RS *studies were classification-based, with little focus on regression tasks. Furthermore, most existing SOC retrieval analyses were conducted outside of Africa. These studies were mostly done using hyperspectral data and TNN frameworks, implying that more research into the effectiveness of DL on different sensors and platforms in estimating SOC is required, especially in Africa. Future SOC studies should consider the use of DL in conjunction with multispectral sensors (e.g. Sentinel and Landsat series) whose unique attributes present greater opportunities for large scale mapping. Consequently, the next chapter will investigate the use and performance of DL-based algorithms in comparison to other TNN and conventional ML models, for large scale SOC stocks mapping. In addition, the study will evaluate the use of Sentinel- 3 data within regional SOC mapping, which is yet to be explored. This will help to assess how DL frameworks perform against other ML models for SOC mapping, whilst ascertaining the viability of using DL and modern multispectral sensors for large scale mapping.*

Chapter Four:

Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data

This chapter is based on;

Odebiri, O., Mutanga, O., & Odindi, J. (2022). Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data. *Geoderma*, *411*, 115695.



Geoderma
Volume 411, 1 April 2022, 115695

Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data

Omosalewa Odebiri, Onisimo Mutanga, John Odindi

https://doi.org/10.1016/j.geoderma.2022.115695

Highlights

- First national scale deep learning-based remote sensing mapping of SOC in South Africa.
- Performance of Deep neural network was compared to other classical machine learning models.
- Results revealed that model type has a significant impact on prediction accuracy.
- Sentinel-3 performed relatively well as a single data source for national scale SOC mapping.
- Sentinel-3's high spectral resolution compliments its low spatial resolution.

Recommended articles

How to estimate soil organic carbon stocks of a...
Geoderma, Volume 411, 2022, Article 115693

Spatial statistical modeling of arsenic accumula...
Geoderma, Volume 411, 2022, Article 115697

Organic matter in aqueous soil extracts: Predict...
Geoderma, Volume 411, 2022, Article 115678

Article Metrics

Citations

Citation Indexes:	1

Captures

Readers:	11

Social Media

Shares, Likes & Comments:	2
Tweets:	1

Abstract:

Mapping of soil organic carbon (SOC) at the regional level is critical for a holistic climate change policy and mitigation of its adverse effects. However, reliable SOC estimates particularly over a large space remains a major challenge due to among others limited sample points, quality of simulation data and the algorithm adopted. Remote sensing (RS) strategies have emerged as a suitable alternative to field and laboratory SOC determination, especially at large spatial extent. The use of Sentinel-3 sensor, the latest of the Sentinel series is minimal and has not been fully developed, despite its impressive attributes that include high spectral-temporal resolution and large coverage. Compared to linear and classical ML models, deep learning (DL) models offer a considerable improvement in data analysis due to their ability to extract more representative features and identify complex spatial patterns associated with big data. Yet, there is paucity in literature on the application of DL-based remote sensing strategies for SOC prediction. Consequently, this study adopted a deep neural network (DNN) to predict SOC at a national scale, using Sentinel-3 image, and compared the results with random forest (RF), support vector machine (SVM) and artificial neural network (ANN) models. The models were trained based on 10-fold cross-validation with 1936 soil samples and 31 predictors. The DNN model generated the best result with a root mean square error (RMSE) of 10.35 t/ha (26% of the mean), followed by RF (RMSE = 11.2 t/ha), ANN (RMSE = 11.6 t/ha) and SVM (RMSE = 13.6 t/ha). The analytical prowess of the DNN, together with its ability to handle big data by learning patterns through a series of hidden layers (6) to draw conclusions, gives it an edge over other classical ML models. The study concluded that the DNN model with Sentinel-3 data is promising and provides an effective framework for continuous national level SOC modelling.

Keywords Deep neural network, Remote sensing, Soil organic carbon, Sentinel-3

4.1. Introduction

Increasing carbon emissions and their effects on different ecosystems have attracted global attention (Wang *et al.*, 2021). Member countries of the Intergovernmental Panel on Climate Change (IPCC) are tasked with continuous quantification and monitoring of carbon emissions. Consequently, carbon pools, including soil organic carbon (SOC), are increasingly attracting research interest as a means to assimilate emitted carbon and mitigate associated adverse impacts (Odebiri *et al.*, 2020b). As the largest terrestrial carbon reservoir, SOC account for about 50% to 80% of the global carbon storage and approximately 2 and 3 times more than the carbon content of the atmosphere and biosphere, respectively (Li *et al.*, 2021; Sahoo *et al.*, 2019). As such, a small change in SOC reserves can significantly affect the global carbon cycle and soil's physical, chemical, and biological properties (Lamichhane *et al.*, 2019). Furthermore, SOC provides valuable information on soil fertility, anion/cation exchange capacity, soil accumulation, soil degradation, water holding capacity and changes in the availability of nutrients that promote vegetation growth (Wang *et al.*, 2018). However, reliable SOC stock mapping remains a big challenge, generally due to; (1) the limited number of available data points, especially at a landscape scale, (2) the type of auxiliary information used in SOC simulation, and (3) the potency and accuracy of interpolation techniques or algorithm adopted (Phachomphon *et al.*, 2010; Angelopoulou *et al.*, 2019).

Advances in remote sensing (RS) has heralded a new era of big data where earth observation satellites are being lunched at a record pace, leading to the availability of a large amount of diverse datasets (Madileng *et al*., 2020; Kumar & Mutanga 2018; Odindi *et al*., 2016). This has opened up a suitable alternative to field and laboratory SOC determination strategies (Mngadi *et al.*, 2020; Hamida *et al*., 2018; Yang *et al*., 2016). Furthermore, Guo *et al*., (2020), notes that when the relationship between traditional soil forming factors and soil properties is weak or negligible, relevant RS data can provide rich supplementary information for a reliable digital soil mapping (DSM). While the aforementioned benefits of RS data to SOC modelling is exciting, its diverse nature, together with very large and ever-growing volumes calls for fast and transferrable analytical techniques for large-scale geospatial information mining in order to maximize its potential (Gupta *et al.*, 2018). Several geostatistical and classical ML models have been investigated to link field-measured SOC to RS metrics (Padarian *et al.,* 2020; Wang *et al.*, 2021). For example, Phachomphon *et al.,* (2010) used various geostatistical techniques to estimate SOC in Laos, using ordinary co-kriging (OCK) producing an $R^2 = 0.42$ accuracy. Similarly, Zhang *et al.,* (2021), used Sentinel-2 and MODIS derived variables to examine the performance of random forest (RF), support vector machine (SVM), and artificial neural network (ANN) in predicting SOC in the northern Songnen plain of China. The RF had the lowest RMSE (0.68 %) and the highest R^2 of all the groups (0.67). While these models have provided acceptable accuracies, they have also shown to be limited in their ability to extract more complicated non-linear abstract elements required for improved predictive models (Wang *et al.,* 2021; Wadoux *et al.,* 2019). Besides, SOC within and between regions exhibit large variability due to their complex blend of organic and inorganic components, hence classical ML models may be unable to accurately predict the behaviour of such a complex phenomenon, especially at large spatial scales (Ma *et al*., 2019; Padarian *et al*., 2019; Kumar *et al*., 2016).

Recently, deep learning (DL) models have gained interest in the RS field because they are well suited in handling big data and its associated heterogeneity (Zhang *et al.,* 2019). DL is a representation-learning technique containing many layers (input, hidden and output) in which information is derived from the lower to higher layers through nonlinear modules (Zhu *et al.,* 2019). DL has proven to be a remarkable improvement to classical ML models, and an exceptionally powerful tool in many fields (Odebiri *et al*., 2021). In contrast to classical ML models, DL models are capable of automatically extracting invariant and abstract features from RS data to improve accuracy. According to Minh *et al.,* (2018), DL architectures can help enhance learning procedures, particularly when relationships between different environmental properties are complex and non-linear. In addition, the inherent multiple hyper-parameters embedded in DL models provides users the opportunity to fine-tune learning procedures in order to improve accuracy (Odebiri *et al.,* 2021). While these DL advantages exist, research on the application of DL-based RS strategies to SOC prediction is lacking and most existing studies are localized with little global impact as DL has only been recently introduced (Odebiri *et al.,* 2021; Padarian *et al*., 2020). To this end, leveraging on the abundance of RS data and the analytical prowess of DL architectures could offer great potential for rapid, continuous and reliable national scale estimates of SOC, thus informing national climate policies and soil management.

In most cases, deep learning (DL) models for SOC mapping is adopted through hyperspectral images (Odebiri *et al*., 2021). For instance, a review of literature by Ma *et al*., (2019) showed that more than 85% of the existing DL-based SOC mapping is performed using hyperspectral data with an average spatial resolution of 2m. However, hyperspectral data remain expensive, especially for applications with wide coverage, and are hard to obtain in many regions, including Africa (Mutanga *et al.,* 2015). Sensors such as the Sentinel series, which are freely available, are excellent alternatives that have yet to be fully explored with DL models. Sentinel sensors have relatively high spatial and spectral resolutions as well as strategically placed bands sensitive to SOC (Odebiri *et al*. 2020a). Although a range of DL studies have been conducted with multispectral data such as Sentinel-2 (Taghizadeh-Mehrjardi *et al.,* 2020) and Landsat 8 (Wu *et al*., 2019), there is little to no exploration of the latest Sentinel-3 data which is critical for national and regional scale mapping (Zhou *et al.*, 2021).

The Sentinel-3 Ocean and Land Colour Instrument (OLCI) from the European Space Agency (ESA) is the latest in the Sentinel series. It comprises four multifunctional satellites (i.e. 3A, 3B, 3C, and 3D). Among them, Sentinel 3A and 3B were launched on February 16, 2016 and April 25, 2018, respectively. Unlike Sentinel-2, which has 13 spectral bands ranging from 443nm to 2190nm wavelengths and a 5-day revisit cycle, Sentinel-3 has better spectral resolution, with 21 spectral bands (400-1020nm) and a shorter revisit time of less than 2 days (Li and Roy, 2017; Kokhanovsky *et al.*, 2019). Compared with Sentinel-2 (290 kilometres), Sentinel-3 has a wider swath width (1270 kilometres), allowing for capture of a large spatial extent at an overpass. However, the former has a better spatial resolution (10 m) than the latter (300 m) and has been successfully used as a single data source for many SOC mapping tasks (Vaudour *et al.,* 2019). Despite the spectral and temporal advantages of Sentinel-3, it is still new and has not be widely used to simulate soil properties, including SOC. Therefore, this study explored a DL approach for SOC modelling across South Africa using Sentinel-3 data. A comparison was also made between the results of the DL method and those of other ML models (RF, ANN, and SVM) commonly used in digital soil mapping.

4.2. Methods and Materials

4.2.1. Soil data

Soil data was obtained from two sources; 1736 points from the International Soil Reference and Information Centre (ISRIC) and the remaining (200) from the Agricultural, Earth and Environmental Sciences Department (SAEES), University of KwaZulu-Natal, South Africa. ISRIC is an independent scientific foundation whose mission is to provide high-quality information on different soil properties (including SOC) at a global scale through cooperation with different countries. The current ISRIC soil database was last updated in 2020 (https://www.isric.org/) and contains more than 150,000 sample points across 173 countries collected at different times (Batjes *et al.*, 2020). Since the SOC carbon content determination methods may vary from country to country (Venter *et al.*, 2021; Hengl *et al.*, 2017), the ISRIC implemented a standardized procedure to make the input soil profile data uniform and available for public use (https://www.isric.org/explore/wosis/accessing-wosis-derived-datasets). The sample points covering South Africa and their equivalent SOC content together with the bulk density were filtered from the ISRIC database in addition to the available data obtained from SAEES. The SOC stock at each point was determined by using the formula: SOC stock (t/h) = SOC concentration x Bulk density x Soil depth (Pearson *et al.,* 2007). In total, 1936 sample points were used for the target variable at a depth of 30 cm.

4.2.2. Image data acquisition

Sentinel 3 OLCI data (Table 4.1) downloaded from the European Space Agency (ESA) between March and April 2021, with less than 10% cloud cover was used in this study. Sentinel-3 is characterized by a large swath width (1270 kilometres), eliminating the time consuming need for numerous downloads and tiling to cover the whole country. Four image tiles were downloaded to cover the entire country. Each of these image tiles were pre-processed for geometric, radiometric and atmospheric correction using the Sentinel Application Platform (SNAP v. 6.0). The ENVI 5.2 software was used to mosaic the pre-processed images into one single tile and the country's shape-file was used to extract its boundary.

Table 4.1. Sentinel-3 data spectral specification

Band No.	Wavelength (nm)	Bandwidth	SNR (at Lref)
1	400	15	2188
2	412.5	10	2061
3	442.5	10	1811
4	490	10	1541
5	510	10	1488
6	560	10	1280
7	620	10	997
8	665	10	883
9	673.75	7.5	707
10	681.25	7.5	745
11	708.25	10	785
12	753.75	7.5	605
13	761.25	2.5	232
14	764.375	3.75	305
15	767.5	2.5	330
16	778.75	15	812
17	865	20	666
18	885	10	395
19	900	10	308
20	940	20	203

21	1020	40	152

Furthermore, different band combinations from the Sentinel-3's twenty-one spectral bands were used to generate relevant spectral vegetation indices used in this study. Literature has demonstrated that vegetation indices can be used to explain the distribution and variability of SOC (Wang *et al*., 2021; Odebiri *et al*., 2020b; Guo *et al*., 2020; Dotto *et al*., 2018). Ten popularly used vegetation indices were generated together with Sentienl-3's 21 spectral bands as independent variables to model SOC. The specification and justification for the use of these indices is summarised on Table 4.2.

Table 4.2. Sentinel-3 derived spectral vegetation indices

Index	Reason for use	Formula	Reference
Normalized Difference Vegetation Index(NDVI)	Calculate vegetation density by separating soil from vegetation and reduce the impact of terrain.	$\frac{B17 - B8}{B17 + B8}$	Rouse (1974)
Soil Adjusted Vegetation Index (SAVI)	Adjusts soil brightness where less vegetation is present	$\frac{B17 - B8}{B17 + B8 + 0.5}(1 + 0.5)$	Huete (1988)
Optimized Soil Adjusted Vegetation Index (OSAVI)	It can adapt to greater soil changes	$(1 + 0.16)\ (B17 - B8)/(B17 + B8 + 0.16)$	Jamalabad and Abkar , (2004)
Modified Soil Adjusted Vegetation Index (MSAVI)	Rectifies areas where soil surface is highly exposed	$\frac{2B17 + 1 - \sqrt{(2B17 + 1)^2 - 8(B17 - B8)}}{2}$	Qi (1994)
Enhanced Vegetation Index (EVI)	Enhance the signal and sensitivity of vegetation in high biomass areas	$2.5 \times \frac{B17 - B8}{(B17 + 6 \times B8 - 7.5 \times B6 + 1)}$	Huete (1999)
Ratio Vegetation Index (RVI)	Indicates amount of Vegetation	$\frac{B17}{B8}$	Baret (1991)
Renormalized Difference Vegetation Index (RDVI)	Linearizes the link between the index and biophysical parameters	$\frac{(B17 - B8)}{(B17 + B8)^1/2}$	Roujean and Breon (1995)
Transformed Vegetation Index (TVI)	Sensitive to and indicative of vegetation	$\sqrt{(NDVI)} + 0.5$	Deering (1975)
Difference Vegetation Index (DVI)	Distinguishes between soil and vegetation	$B17 - B8$	Richardson (1977)
Green Normalized Difference Vegetation Index (GNDVI)	Receptive to vegetation and plant chlorophyll content	$\frac{B17 - B6}{B17 + B6}$	Gitelson (1998)

4.2.3. Analytic models

A regression analysis for SOC retrieval was performed based on a deep neural network (DNN) framework. DNN's performance was then compared to three commonly used classical machine learning (ML) algorithms in digital soil mapping (i.e. random forest -RF, support vector machine -SVM and artificial neural network -ANN) (Odebiri *et al.,* 2021; Zhou *et al.,* 2020). As non-parametric models, these algorithms (SVM, RF and ANN) have been extensively used in literature to model non-linear relationships between SOC and various covariate predictors (Somarathna *et al.,* 2017). In addition, each of the selected model in this study contains hyper-parameters in its structure, which when properly configured, could have a major effect on model performance (Chen *et al.,* 2019). Hence, we conducted hyper-parameter optimization using a random search technique and a 10-fold cross-validation with a train/validation/test split for each model to produce the best possible result. The analysis was conducted using the python programming language (version 3.8) within the jupyter notebook. A brief description of the four algorithms together with a summary of the final selected hyper-parameters is presented in Table 4.3.

Table 4.3. Algorithms defined hyper-parameters

Algorithm	Hyper-parameters	Parameter as used	Parameter Description
DNN	Hidden layer	3-20	Number of hidden layers
	Node Size	100–500	Number of nodes (neurons) in the hidden layers
	Network weight	uniform/he-normal	Initialized weights of the layers from input to output
	initialization		Adjust the weights of the network
	Learning rate	0.001–0.05	Neurons randomly dropped during training to reduce
	Dropout	0.2–0.3	overfitting
	regularization	100-1000	Number of training iteration
	Epochs		
RF	Mtry	1–20	Input variables number
	Ntree	100–1000	Number of trees
	Node size	1	
SVM	Kernel type	RBF	Radial basis function
	C	0.01–100	Penalty specification
	σ	0.01–100	Bandwidth specification
ANN	Decay	0.001–0.05	Weight regulator
	Size	2-10	Nodes or units in the hidden layer
	Epochs	10-100	Number of training iteration

4.2.3.1. Deep Neural Networks (DNN)

The use of DNN architectures in different applications has increased in the last few years, including, earth observation, image recognition, automated driving, and speech recognition

(Pudełko & Chodak., 2020; Dong *et al.,* 2018; Novoa *et al*., 2018;). Essentially, the DNN is an effective and reliable approximate model that provides information about the complex relationships between the target and explanatory variables (Wang *et al*., 2020). Between the input, the hidden and the output layers are many neurons, such that neurons in one layer are linked to the neurons in the next layer until the final predicted (output) neuron. DNN generally uses the multilayer perceptron (MLP) architecture but differs through its many incorporated hidden layers and hyper-parameters (see Table 4.3) that gives it an edge over other classical ML models (Taghizadeh-Mehrjardi *et al.,* 2020). However, care must be taken when training the model to avoid overfitting (Liu *et al*., 2018). In such cases, a dropout regularization technique can be applied to each of the hidden layers, whose neurons are subsequently averaged for prediction (Taghizadeh-Mehrjardi *et al*., 2020). A simple mathematical representation of DNN for a given input layer (vector X) with a L hidden layer and an output layer (vector Y) can be represented as follows (Wang *et al*., 2020);

$$\mathrm{Z}^1 = \sigma_1(\mathrm{W}^1\mathrm{X} + \mathrm{b}^1), \mathrm{Z}^2 = \sigma_2(\mathrm{W}^2 z^1 + \mathrm{b}^2)$$
$$\mathrm{Z}^\mathrm{L} = \sigma_\mathrm{L}(\mathrm{w}^\mathrm{L} z^{L-1} + \mathrm{b}^\mathrm{L}),\ \ \mathrm{Y} = \ \mathrm{W}^{\mathrm{l}+1}\mathrm{Z}^\mathrm{L} + \mathrm{b}^\mathrm{L}, \theta = [\mathrm{W}^\mathrm{i}, \mathrm{b}^\mathrm{i}]_{\mathrm{i}=1}^{\mathrm{L}+1} \quad (1)$$

Where W^i and b^i are weights and biases of the i^th layer respectively. $L + 1$ indicate the output layer (i.e. Y = N (X; *Ɵ*)) and σ^{i} is the activation function which can either be sigmoid, softmax, rectified linear unit (ReLU), hyperbolic tangent (Tanh) among others.

The loss function *L* of the output and input variable can be examined using the mean squared error (MSE) represented as;

$$\mathrm{MSE}_{\mathrm{DATA}} = \mathrm{L}(\theta) = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} |\mathrm{NN}(\mathrm{X}_\mathrm{i};\ \theta) - \mathrm{Y}_\mathrm{i}|^2 \quad (2)$$

Where *N* is the sum total of the labelled data, and L is the loss function which can be minimized using an optimization algorithm (Wang *et al*., 2020). Figure 4.1 shows the schematic representation of the DNN architecture. A total of 31 input variables generated from Sentinel-3 OLCI were used to build the model in this study. The complete dataset was partitioned into ten equal parts and used for training and testing in a sequential order. To ensure that each data point was utilized as validation at least once, the DNN was calibrated ten times. After repeated adjustments using a random search optimization with 10-fold cross-validation, 500 epochs, six hidden layers, the “adam” optimizer and the ReLU activation function were used for the best output.

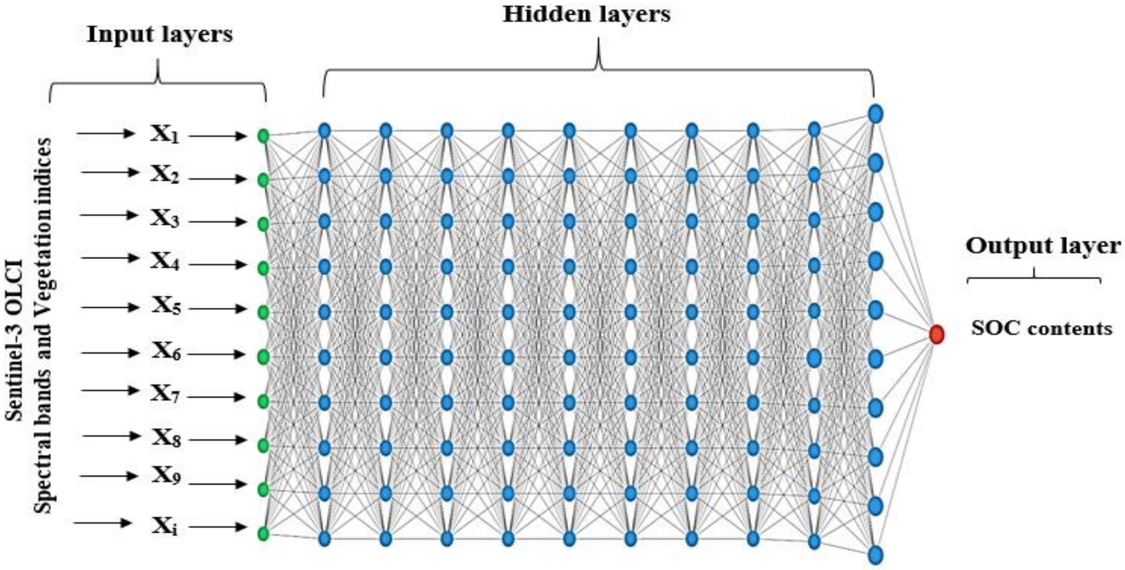


Figure 4.1. A graphical illustration of the deep neural network(DNN) framework

4.2.3.2. Random forest (RF)

Random forest is a popular and widely adopted machine leaning algorithms for regression and classification task (Sibanda *et al.,* 2021). It is a non- parametric tree-based ensemble algorithm that is capable of handling both small and big data (Rasaei and Bogaert, 2019). During training, RF uses an average of the large decision trees combined with a unique bootstrap sampling to obtain predicted results (Odebiri *et al*., 2020a). RF model also contains an important function (Gini impurity) within its structure that helps the user to determine how each predictor variable contributes to the overall success of the model. (Breiman, 2001). The hyper-parameters including *ntree*, *mtry* and the *node-size* of the RF can be optimised to produce better predictive results (Table 4.3). In addition, bootstrapping reduces the effect of overfitting in RF and allows accurate estimation of errors from the out-of-bag (OOB) samples (Were *et al.*, 2015). The OOB mean squared error (MSE_{OOB}) is expressed as the summation of the predictions of all trees as follows (Zhou *et al*., 2020);

$$\mathrm{MSE_{OOB}} = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i^{OOB})^2 \tag{3}$$

Where *n* is the number of observations and $\hat{Z}_i^{OOB}$ is the OOB prediction for observation z_i. The RF model in this study was calibrated by splitting the data into 10 equal parts used for training and testing sequentially with a defined set of hyper-parameter as shown in Table 4.3.

4.2.3.3. Support vector machine (SVM)

Like RF, the SVM is also widely used for regression and classification tasks in remote sensing applications (Forkuor *et al.*, 2017). SVM is a non-parametric model that works by constructing a set of hyperplanes in an infinite dimensional space (hyperspace) with the help of a kernel function (Were *et al.*, 2015). The use and choice of the kernel function including sigmoid, polynomial, radial basis function (RBF), and linear, are critical to overall SVM performance (Jeong *et al.*, 2017). Given a set of predictor variables (x) to simulate the target variable (y), the basic principle of SVM can be simply expressed as follows (Guo *et al.*, 2020);

$$\min_{w,b}\frac{1}{2}\,||w||^2,$$

$$\text{s.t.}\, y_i(w^T x_i + b)\ \geq 1, i = 1,2,3, \ldots\ldots m. \tag{4}$$

Where x_i and y_i are the predictors and the target variable respectively; w and b are the vector of the hyperplane and bias.

The RBF kernel function was selected for this study because it is commonly used in soil mapping and performs better than other kernels. (Zhou *et al.*, 2020; Guo *et al.*, 2020; Keskin *et al.*, 2019). We also defined and optimized two other important RBF parameters (i.e. penalty — cost, sigma — kernel width) using the random search technique within the Python 3.8 API.

$$K\big(x_i, x_j\big) = \exp(-\sigma\big|x_i +\ x_j\big||^2) \tag{5}$$

Where K represents the kernel function defined by the user (RBF), x indicates the input vector, and σ is the sigma (Jeong *et al.*, 2017). The data partitioning for the SVM analysis was similar to the RF model as specified above.

4.2.3.4. Artificial neural network (ANN)

ANN is a classical neural network mainly used for classification and regression analysis (Wang *et al.*, 2021). Its unique and vigorous data-modelling ability helps it to detect patterns and draw result, thus can be applied to simulating complex soil properties such as SOC (Xu *et al.*, 2020). The ANN consist of input, hidden and output layer (Wei *et al.*, 2021; Falahatkar *et al.*, 2016). The input layer (x_i) receives the signal (i.e. data), following which, the hidden layer performs a weighted linear combination of multiple independent variables and stores the model parameters such as weight and bias (Gruszczyński, 2019). Before making the final prediction (output layer y_i), a non-linear activation function is applied to limit the influence of outliers (Chen *et al.*, 2019). The ReLU activation function was used with a single hidden layer in this study (Kuang *et al.,* 2015). The ANN in this study was made by setting the learning rate, the number of nodes at the hidden layer and epochs to 0.001–0.05, 2–10, and 10-100 respectively as shown in Table 4.3.

4.2.4. Model evaluation metrics

The fitting and generalization of the models built in this study were evaluated using three metrics namely: root mean squared error (RMSE), coefficient of determination (R^2) and Lin's concordance correlation coefficient (LCCC). These metrics are expressed as;

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_{O,i} - X_{P,i}\right)^2}{n}} \tag{6}$$

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} (X_O - p)^2}{\sum_{i=1}^{n} (X_O - O')^2}\right] \tag{7}$$

$$\mathrm{LCCC} = \frac{2\, r\, \sigma_O \sigma_P}{\sigma_p^2 + \sigma_p^2 + [O' - P']^2} \tag{8}$$

Where *n* signify number of observations, X_O and X_P are the measured and predicted SOC value. *O'* and *P'* represents averages of the measured and predicted SOC, while $\boldsymbol{\sigma_o}$ and $\boldsymbol{\sigma_p}$ are the respective variance of the measured and predicted value. Furthermore, to avoid sampling bias, a 10-fold cross-validation was performed by dividing the data into 10 equal sets and passed sequentially into both calibration and validation datasets, so that each set was used at least once. Generally, a best-fitted model is defined by a higher R^2 and LCCC, along with a lower RMSE.

Furthermore, the importance of predictor variables was evaluated to assess their contribution to the overall performance of each model. Whereas the RF model has a function (Gini impurity) in its structure to measure the importance of variables, other algorithms do not. Specifically, deep learning (DL) models often fail to achieve interpretability because they cannot quantify the importance of variables for regression or classification tasks, which is why they are called a black box (Padarian *et al.,* 2019). To this end, several techniques have recently been proposed to help users interpret DL predictions (Pentos, 2016). One of the methods is the SHApely Additive exPlanations (SHAP) used in this study. The working principle of SHAP is to assign a specific average value to each variable to show the magnitude of their impact on the model output. SHAP possesses special function for every type of ML model, including "DeepExplainer" for DL and ANN, "TreeExplainer" for any tree-based model (RF), and "KernelExplainer" for other types of models (SVM). The advantages of SHAP over other strategies include, but are not limited to global and local interpretability (for more details on SHAP, see Padarian *et al.,* 2020, and Lundberg and Lee, 2017). A comparison between the SHAP value and the RF Gini impurity yielded the same result, therefore, it was used as a unified method for feature importance measure in this study.

4.2.5. Uncertainty quantification

In addition to evaluating the performance of a model, it is a good practice to quantify its uncertainty (Abdar *et al.,* 2021). In decision-oriented research, characterizing uncertainty and evaluating the robustness of research conclusions are crucial to achieving the quality and credibility of the analysis (Hamel and Bryant, 2017). Uncertainty quantification provides the upper and lower bounds for the estimated output variables (Abdar *et al.,* 2021). In this study, the upper and the lower bounds for the SOC maps generated by each model were determined using the common ±1.64 standard deviation (SD) with a corresponding significance level of 90% confidence interval (C1) (Minasny *et al.,* 2016). This was done using a 10-fold cross-validation under the assumption that the four models follow a normal distribution for every raster cell (Emadi *et al*., 2020). Subsequently, the 5th and 95th percentiles together with the predicted mean value of each pixel was retrieved. Finally, a spatial distribution map for the calculated mean, lower (5%) and higher (95%) confidence interval was generated for the four models.

4.3. Results

4.3.1. Summary statistics

Statistical analysis was conducted to describe the target SOC data (1936 samples) for the predictions. The data range between 5.3 t/ha and 149 t/ha, with average and standard deviations of 39.8 t/ha and 17.3 t/ha, respectively. The coefficient of variation (43%) is calculated as the ratio between standard deviation and mean, indicating relatively high variation within the SOC data. The data also showed strong skewness (1.9) and kurtosis (5.2), which nullified the normal distribution rule (Hair *et al.,* 2016). In order to improve the normal distribution, we applied a natural logarithm transformation to the SOC data, resulting in skewness and kurtosis of 0.41 and 0.68, respectively. The converted SOC data was inversely transformed to restore the data to its original scale after prediction analysis.

4.3.2. Evaluation and performances of models

The average results for the four models (i.e. DNN, RF, ANN, and SVM) using 10-fold cross-validation are presented in Table 4.4. Among the four trained models, the DNN model showed the strongest robustness in predicting SOC variability, indicating a value of 10. 35 t/ha (26% of the mean) for RMSE, 67.3 for R^2, and 84.3 for LCCC. Other model's performance in order of ranking were RF with an RMSE score of 11.2 t/ha (28% of the mean), 64.7 for R^2 and 80.5 for LCCC; the ANN with scores of 11.6 t/ha (29% of the mean), 63.4 and 79.6, for RMSE, R^2 and LCCC respectively; and the SVM was the least robust, denoting an RMSE score of 13.6 t/ha (34% of the mean), 58 for R^2 and 77.5 for LCCC, respectively. Figure 4.2 shows the correlation between the observed and the estimated SOC for the constructed models.

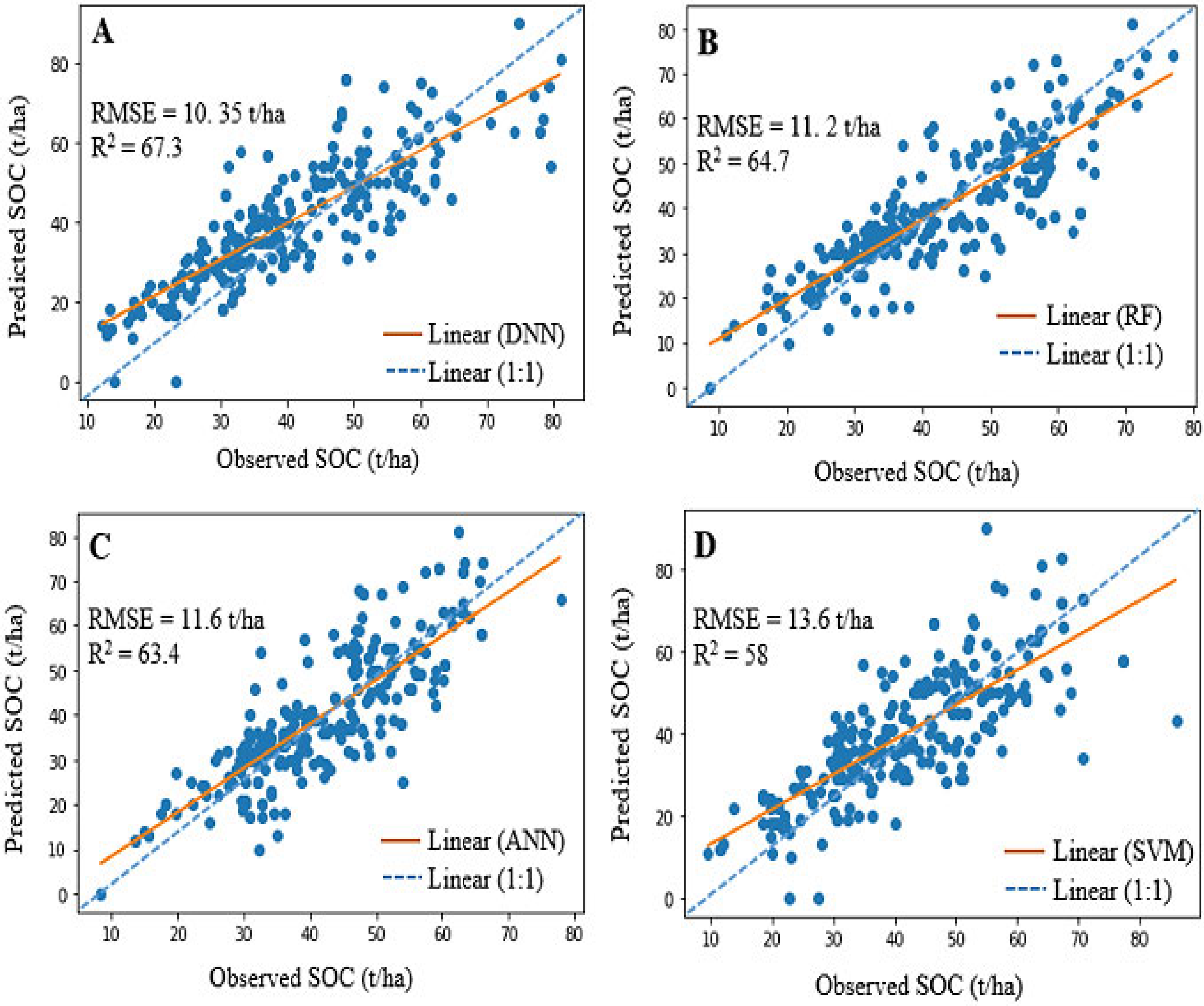


Figure 4.2. Predicted against observed soil organic carbon (SOC) using four models: (A) DNN = deep neural network, (B) RF = random forest, (C) ANN = artificial neural network, (D) SVM = support vector machine

Table 4.4. Results of deep neural networks (DNN), random forests (RF), artificial neural networks (ANN), and support vector machines (SVM); RMSE: root mean square error; R^2: coefficient of determination; LCCC: lin's concordance correlation coefficient (mean ± standard deviation).

Model	RMSE (t/ha)	R^2	LCCC
DNN	10. 35 ± 5.05	67.3 ± 5.01	84.3 ± 6.05
RF	11.2 ± 4.03	64.7 ± 5.03	80.5 ± 6.02
ANN	11.6 ± 6.02	63.4 ± 5.0	79.6 ± 6.02
SVM	13.6 ± 5.07	58 ± 5.03	77.5 ± 7.08

4.3.3. Variable importance assessment of models

Figure 4.3 shows the relative importance and contribution of the top twenty performing covariates using the SHAP approach. The figure shows that the best explanatory variable for DNN and ANN models was Band 8 (665 nm), while NDVI was the most important variable in RF and SVM models. Except for the SVM model, the importance ranking of variables in other models was relatively uniform. For example, the first five predictors of the DNN model, including B8, NDVI, EVI, B11, and RVI, were the same for RF and ANN, with a slight difference in their rankings. A closer look at the figures shows that NDVI and Band 8 were among the top three in all models. In addition, the influence of the first five variables in the best model (DNN) were higher when compared to other models.

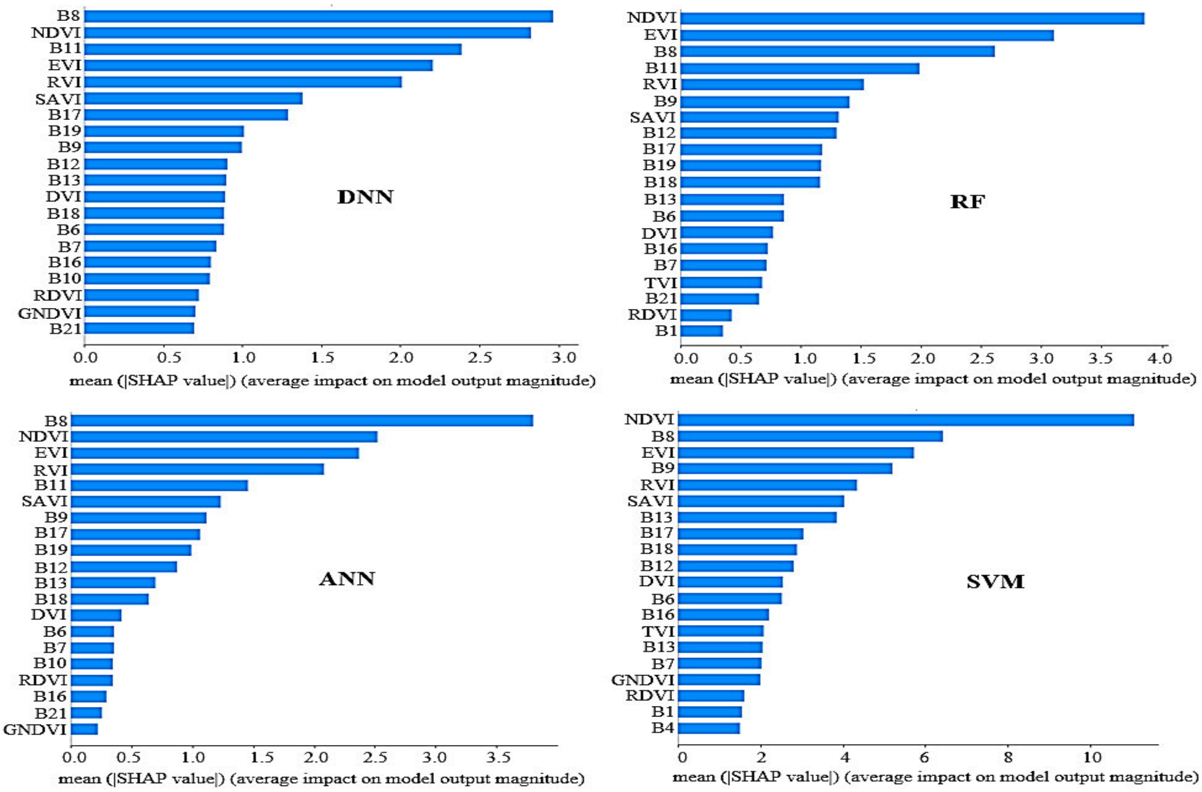


Figure 4.3. Importance ranking of variables used for the simulation of soil organic carbon across South Africa. DNN: deep neural network, RF: random forest, ANN: artificial neural network, SVM: support vector machine

4.3.4. Spatial estimation of SOC and uncertainty quantification

Figure 4.4 and Table 4.5 show the spatial distribution and uncertainty quantification of predicted SOC for South Africa at the upper limit (95%), the mean and the lower limit (5%), calculated at 90% CI. The maps generated by each model (Figure 4.4) show a significant level of agreement from the upper limit to the lower limit. The concentration of SOC is higher at densely vegetated areas compared to areas with sparse or no vegetation. For instance, the north-eastern and south-eastern part with darker colours are dominated by dense natural woodlands and plantations. This is not surprising, because the most important variables (B8, NDVI, B11, EVI and RVI) to the overall performance of the models are all sensitive to vegetation. Additionally, Table 4.5 shows the percentage of observation that falls within the defined CI (i.e. 5 and 95%) for the four models, thus depicting their uncertainty. In theory, 90% of the observations should fall within the specified range (Minansny *et al.*, 2016). The DNN model produced the most reliable uncertainty, with 88% of the observations within the defined CI, followed by RF and ANN with similar uncertainty at about ~79%. The SVM also produced a relative uncertainty to the other models at ~69%. While the models depict a good measure of uncertainty, they may not represent the actual situation on ground, because the quantification of uncertainty in this study is based on the parameters of the models and not on the spatial uncertainty of the data.

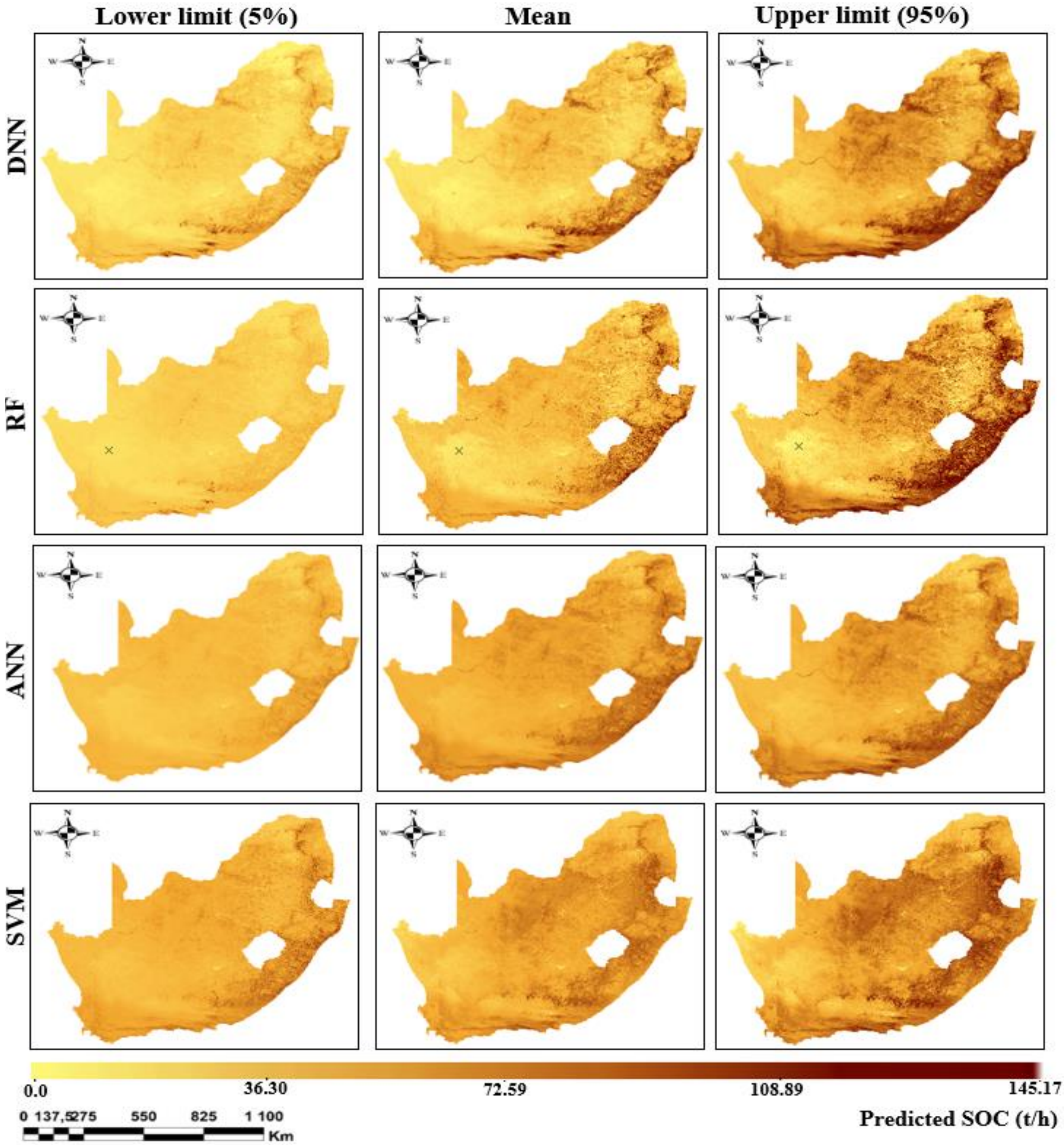


Figure 4.4 The spatial distribution of soil organic carbon (SOC) for deep neural network (DNN), random forest (RF), artificial neural network (ANN) and support vector machine (SVM) at lower limit (5%), mean, and upper limit (95%) using Sentinel 3 OLCI data

Table 4.5. Number and percentage of SOC samples within the defined confidence interval (90%)

Models	Observations	Number of Observations			Expressed in percentage (%)		
		Within CI	Outside CI		Within CI	Outside CI	
		5 — 95%	<5	>95%	5 — 95%	<5	>95%
DNN	1936	1705	136	95	88.07	7	4.9
RF	1936	1541	217	178	79.6	11.2	9.2
ANN	1936	1528	161	247	78.9	8.35	12.75
SVM	1936	1338	301	297	69.1	15.56	15.34

4.4. Discussion

The easy acquisition and rapid growth of various remote sensing data based on spectral reflectance has led to an increase in studies on landscape characteristics, including SOC (Wang *et al.*, 2021). Powerful data driven algorithms such as deep learning (DL) are increasingly being adopted in the remote sensing field due to the proliferation of big data (Ma *et al.,* 2019). Compared to classical ML models, DL adapts to environmental changes and improves models based on continuous feedback (Minh *et al.,* 2018). In addition, DL is facilitated by neural networks, multiple hyper-parameters and multiple-layer architecture capable of improving accuracy (Zhu *et al.,* 2019). It is an advanced form of machine learning (ML) which collects data, learns from it, and optimizes the model (Li *et al.,* 2021). In this study, we examined a deep learning approach (DNN) to predict the spatial distribution of SOC at a national scale using Sentinel-3 OLCI and its derivatives. The result and performance of the DNN was compared to other popular ML models used in digital soil mapping (i.e. RF, ANN, and SVM).

The DNN model produced the best results for SOC distribution with a considerable level of accuracy when comapred to the other models (Table 4.4). This is consistent with existing literature. A recent study by Emadi *et al.*, (2020) showed the superiority of DNN over other models such as RF, ANN and SVM in mapping SOC variabilty across the Mazandaran province of northern Iran with R^2 of 65, 58, 55, and 53 respectively. Similarly, Taghizadeh-Mehrjardi *et al.*, (2020) developed a seven hidden layer DNN model to predict SOC content at six standard depths across two contrasting sites in central and northern Iran, and obtained a better result at all depths compared to other machine learning models. In addition, the results ($R^2 = 67.3$) obtained in this study exhibited a better performance compared to other recent national-scale SOC studies by Venter *et al.*, (2021), who used RF model, and Schultze & Schutte (2020) who calculated SOC % using area-weighting, with R^2 score of 65.9 and 20.3

respectively. It is important to note that the soil sample points used in both studies were three to four times higher than this study, yet the DNN model still produced a better accuracy. The edge of the DNN model over other models is based on its ability to learn and extract more representative features from the SOC data through its many hidden layers and neurons. Each neuron in the network represents an aspect of the data, and together they provide a complete representation of the data. The hidden layers (in our case 6) are weighted to indicate the strength of their relation to output (i.e. SOC), and as the model develops during training, the weights are adjusted to improve accuracy. In addition, the DNN can accurately approximate the complicated non-linear relationship between SOC and covariates, thus capturing the potential association between them (Yuan *et al.,* 2020). Considering the spatial uncertainty of the target SOC data due to multiple-scales of variation as well as the different sampling sources and time of collection, the DNN model (6 hidden layers) was more precise than other methods, indicating its robustness in complex data modelling.

This study also evaluated the importance of predictors used to explain the variability of SOC. The SHAP value technique with a special ability to reveal the importance of predictors to any machine or deep learning model was adopted (see Section 2.4). The top five variables in the DNN model were Band 8 (B8 = 665nm), NDVI, Band 11 (B11 = 708.25nm), EVI, and RVI. Interestingly, these variables were also significant in other models occupying the first six positions in rankings (Figure 4.3). The most important variable (B8) of the DNN model falls within the visible red spectrum (625-700nm). Many studies have demonstrated the sensitivity of the red band to SOC (Gholizadeh *et al.,* 2018; Mondal *et al.,* 2017). For instance, Odebiri *et al.,* (2020) recently highlighted the importance of the red band of Landsat-8 data to estimate SOC content within the same study area (South Africa). According to their study, the red band region was highly responsive to vegetation attributes such as chlorophyll content, which provides important information on the physiological state of vegetation related to SOC. In addition, NDVI, EVI, and RVI also improve vegetation detection and sensitivity in high biomass regions, indicating density and distribution of vegetation, which in turn inform SOC variability (Kumar *et al.,* 2016; Matsushita *et al.,* 2007). The B11 covers the strategic regions of the vegetation-sensitive electromagnetic spectrum (red-edge). According to Mngadi *et al.,* (2019), the red-edge region provides information on a wide range of vegetation attributes, including biomass, canopy structure and chlorophyll content. Previous studies (e.g. Zhang *et al.,* 2019; Aryal *et al.,* 2017; Nabiollahi *et al.,* 2019; Forkuor *et al.,* 2017) have reported that SOC concentration is highly dependent on vegetation intensity and residues. The same applies to our SOC model (Figure 4.4), as the vegetation density and SOC concentration were positively correlated, supporting the hypothesis that vegetation may be utilized as a surrogate for SOC estimation since both respond to the same physical and environmental triggers (Bhunia *et al.,* 2017).

Overall, from the output maps, SOC is more prevalent in the north and south-east of the country, while the desert, arid and semi-arid west has lower concentration. This is because areas of higher SOC concentration are dominated by dense natural and exotic plantations characterized by high canopy coverage and height ranging from 10-75% and 2.5-6 meters, respectively (SANLC, 2020). Densely vegetated areas are capable of promoting accelerated soil metabolism through continuous litterfall and dead matter, resulting in additional SOC accumulation (Muchena, 2017). Furthermore, the densely vegetated areas of the country receive higher rainfall (> 600 mm) than the desert and arid west (< 300 mm). Rainfall influences soil moisture, hydrological processes (including surface runoff and groundwater infiltration), vegetation density, and decomposition, which supports SOC sequestration (Zhou *et al.,* 2008; O'Brien *et al.,* 2010; Chen *et al.,* 2015). As an example, the northern cape province of the country (Figure 1.1— study area map in chapter 1), which has the lowest concentration of SOC (Figure 4.4), is characterized by low erratic long-term annual rainfall with less than 175 mm (Paterson 2014). Likewise, the soils of this area are largely dominated by very shallow sandy soils with a low water retention capacity and wind-blown sands (dunes), leading to a low soil organic carbon accumulation.

The Sentinel-3 data performed well in a nationwide SOC modelling. Sentinel-3 acquires data within the visible (VIS) to near infrared (NIR) wavelength region of the electromagnetic spectrum (400 to 1020 nm). These VIS-NIR wavelength region provides critical reflectance information on SOC and considered the most sensitive region to determine SOC content (Lin *et al*., 2020; Bilgili *et al*., 2010). Whereas its spatial resolution is low (300m), the short revisit period and large coverage provide comprehensive information for SOC estimation at country level (Li *et al.*, 2021). A few studies have examined the capability of Sentinel-3 data to estimate SOC. For instance, Lin *et al*., (2020) compared the performance of Sentinel-3 and 2 to predict SOC content in China and achieved different accuracies of R^2 = 55 and 59, respectively. The study concluded that the multiple spectral bands (21) of Sentinel-3 can complement its low spatial resolution, resulting in improved SOC prediction. In addition, according to the importance ranking of the SHAP value approach used in this study, the following Sentinel-3 bands were important for SOC prediction; B8(665nm) B11(708.25nm), B17(865nm), B19(900nm), B9(673.75), B12(753.75nm), B13(761.25nm), B18(885nm), B6(560nm), and B7(620nm), respectively. We recommend that these band regions be noted for future studies.

4.5. Conclusion

In this study, a national SOC estimation was conducted using a deep learning (DNN) approach and Sentinel-3 OLCI data. The DNN model performed better than RF, ANN and SVM, indicating its ability to extract more abstract features from data, thus, increasing accuracy. We also adopted the SHAP technique that aided in unveiling the lack of interpretability associated with deep learning models. Band 8, NDVI, Band 11, EVI and RVI were the most important variables in the DNN model for determining SOC variability in the study area. However, care must be taken when building the DNN model to avoid overfitting; a phenomenon common to deep learning frameworks due to many hyper-parameters. In addition, this study established the capability of Sentinel-3 data to predict SOC at a regional/country scale. This is important to evaluate its competitiveness in reference to other commonly used sensors like Landsat and MODIS, thus affording end-users an opportunity to choose the appropriate image suitable at different mapping scales. This study is the first developed DL-based SOC mapping in South Africa and provides a valuable framework for improved national carbon accounting. However, cost and sample size must be put into perspective because DL models require a lot of high computing power as well as big data.

4.6. Summary

This chapter presented a novel cost-effective framework for DL-based national scale SOC stock distribution mapping in South Africa, using Sentinel-3 data and a DNN architecture with six hidden layers. The DNN model outperformed other popular models such as RF, SVM and ANN, as well as prior existing South African SOC models. Using the SHAP approach, this chapter also addressed the absence of interpretability constraints that are usually associated with DL frameworks. This allowed the most essential explanatory factors in the DNN model, such as sensitive Sentinel-3 spectral segments and vegetation indices, to be revealed — which are critical for mapping SOC. Having established the superiority of DL models over other TNN and classical ML models, as well as the feasibility of using multispectral sensors for national SOC modelling, it is critical that future research now begin to focus on real-world SOC stock distribution and the influence of different environments on SOC accumulation and sequestration. To that purpose, the next chapter will use Deep Neural Networks (DNN) and

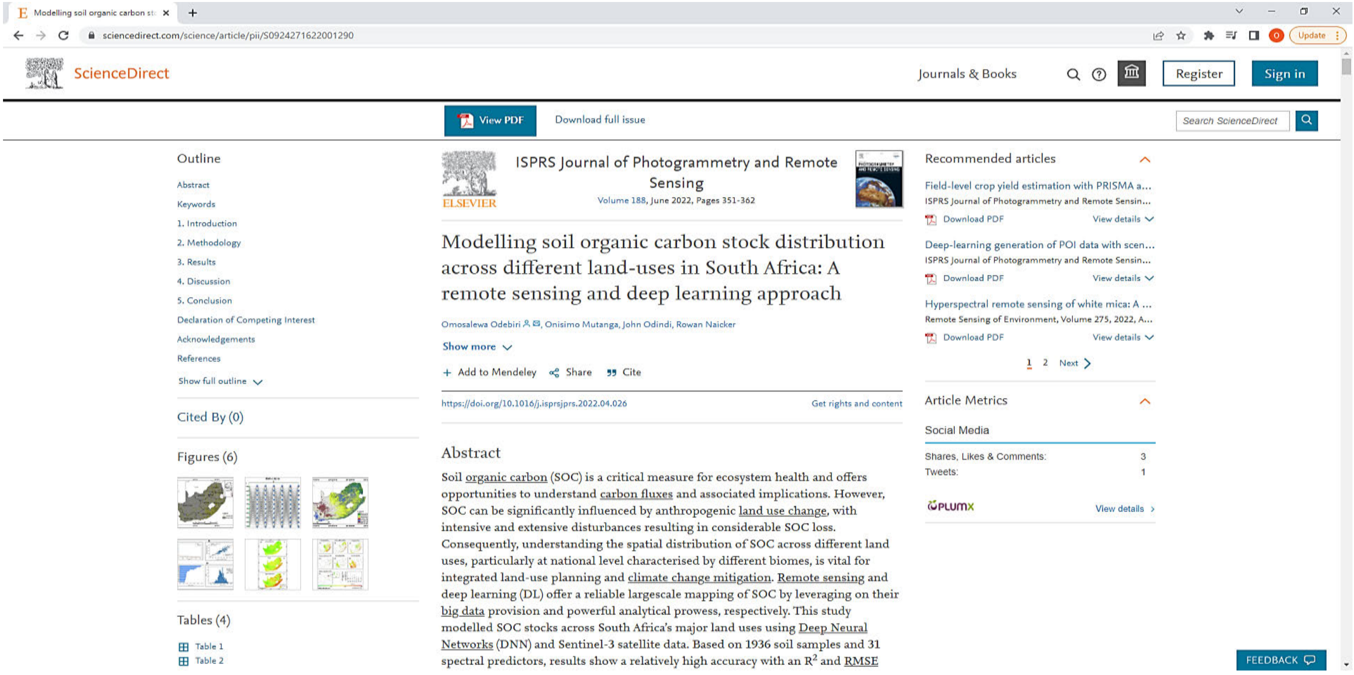
Sentinel-3 satellite data to explore SOC stock distribution across South Africa's main land uses. This chapter will examine the impact of anthropogenic land use change on SOC. Such knowledge will aid in the improvement of climate change mitigation activities, policies, and management practices, as well as provide the government and policymakers with a focus for planned interventions.

Chapter Five:

Modelling soil organic carbon stock distribution across different land-uses in South Africa: A remote sensing and deep learning approach

This chapter is based on;

Odebiri, O., Mutanga, O., Odindi, J., & Naicker, R. (2022). Modelling soil organic carbon stock distribution across different land-uses in South Africa: A remote sensing and deep learning approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, *188*, 351-362.



ISPRS Journal of Photogrammetry and Remote Sensing
Volume 188, June 2022, Pages 351-362

Modelling soil organic carbon stock distribution across different land-uses in South Africa: A remote sensing and deep learning approach

Omosalewa Odebiri, Onisimo Mutanga, John Odindi, Rowan Naicker

https://doi.org/10.1016/j.isprsjprs.2022.04.026

Abstract

Soil organic carbon (SOC) is a critical measure for ecosystem health and offers opportunities to understand carbon fluxes and associated implications. However, SOC can be significantly influenced by anthropogenic land use change, with intensive and extensive disturbances resulting in considerable SOC loss. Consequently, understanding the spatial distribution of SOC across different land uses, particularly at national level characterised by different biomes, is vital for integrated land-use planning and climate change mitigation. Remote sensing and deep learning (DL) offer a reliable largescale mapping of SOC by leveraging on their big data provision and powerful analytical prowess, respectively. This study modelled SOC stocks across South Africa's major land uses using Deep Neural Networks (DNN) and Sentinel-3 satellite data. Based on 1936 soil samples and 31 spectral predictors, results show a relatively high accuracy with an R^2 and RMSE

Recommended articles

Field-level crop yield estimation with PRISMA a...
ISPRS Journal of Photogrammetry and Remote Sensin...
Download PDF View details

Deep-learning generation of POI data with scen...
ISPRS Journal of Photogrammetry and Remote Sensin...
Download PDF View details

Hyperspectral remote sensing of white mica: A ...
Remote Sensing of Environment, Volume 275, 2022, A...
Download PDF View details

1 2 Next

Article Metrics

Social Media

Shares, Likes & Comments:	3
Tweets:	1

PLUMX View details

Abstract:

Soil organic carbon (SOC) is a critical measure for ecosystem health and offers opportunities to understand carbon fluxes and associated implications. However, SOC can be significantly influenced by anthropogenic land use change, with intensive and extensive disturbances resulting in considerable SOC loss. Consequently, understanding the spatial distribution of SOC across different land uses, particularly at national level characterized by different biomes, is vital for integrated land-use planning and climate change mitigation. Remote sensing and deep learning (DL) offer a reliable largescale mapping of SOC by leveraging on their big data provision and powerful analytical prowess, respectively. This study modelled SOC stocks across South Africa's major land uses using Deep Neural Networks (DNN) and Sentinel-3 satellite data. Based on 1936 soil samples and 31 spectral predictors, results show a relatively high accuracy with an R^2 and RMSE value of 0.685 and 10.15 t/h (26% of the mean), respectively. From the seven land uses evaluated, grasslands (31.36%) contributed the most to the overall SOC stocks while urban vegetation (0.04%) contributed the least. Moreover, although SOC stock was found to be relatively proportional to land coverage, commercial (46.06 t/h) and natural (44.34 t/h) forests showed a higher carbon sequestration capacity. These findings provide an important guideline to managing SOC stocks in South Africa, useful in climate change mitigation through sustainable land-use practices. Whereas landscape restoration, and other relevant interventions are encouraged to improve SOC storage, care must be taken within land-use decision making to maintain an appropriate balance between carbon sequestration, biodiversity, and general ecosystem functions.

Keywords: Soil organic carbon, Land use, Land-use planning, Remote sensing, Deep learning, Sentinel-3

5.1. Introduction

Soil organic carbon (SOC) is the main component of soil carbon storage and plays a key role in maintaining ecosystem services, including food production, water supply, soil fertility and climate change mitigation (Jiao *et al.,* 2020). In terrestrial ecosystems, SOC is considered the largest carbon pool, containing three and four times as much carbon that is stored in the atmosphere and biotic pools, respectively (Kenye *et al.,* 2019; Odebiri *et al.,* 2020b). Globally, more than 1,500 Petagrams (Pg) of carbon is stored in the soil as organic carbon (Ghimire *et al.,* 2018). However, small changes in this reserve, caused by the effects of anthropogenic land use change, may significantly alter the global carbon cycle (Sainepo *et al.,* 2018). This may not only deteriorate soil quality, but also emit greenhouse gases (GHG) into the atmosphere, thereby accelerating climate change (Lamichhane *et al.,* 2019; Swanepoel *et al.,* 2016). Generally, changes in land use account for 12-20% of human-induced carbon emissions and are expected to remain one of the largest source of greenhouse gases (Amanuel *et al.,* 2018). According to Sanderman *et al.,* (2017), over the last 12,000 years, the conversion of natural land to cultivation has resulted in approximate loss of 116 Pg and 22 Pg of carbon from grasslands and savanna ecosystems, respectively. Hence, changes in SOC due to changes in land use have received global attention as a key issue for climate change mitigation, agricultural management, ecosystem restoration and environmental conservation (Jiao *et al.,* 2020; Ou *et al*., 2017).

In South Africa, majority of the terrestrial carbon pool is comprised of SOC stocks, with an estimated average sink of 6,396 gC/m^2 and a net primary production of 186 gC/m^2 (Department of Environmental Affairs, 2017). Nevertheless, to cope with agricultural and housing need for the county's rapidly growing population, natural vegetation has been transformed into agricultural and residential land uses, leading to significant SOC losses (Schulze and Scuttle, 2020). For instance, close to 20% of South Africa's natural land cover has been altered by cultivation, land degradation and urbanization (Venter *et al.,* 2017). Griscom *et al.,* (2017), also noted that since 10 000 BC, about 2 Pg of South Africa's carbon has been lost through agricultural activities, while Du Preez *et al.,* (2011) found that perpetual cultivation, especially topsoil disturbances, may reduce soil carbon stocks by approximately 45%. Similarly, Swanepoel *et al.,* (2016), highlighted that agriculture has significantly reduced SOC reserves in Southern Africa by 25-53%.

Recent studies have investigated the impact of land use on SOC stocks in South Africa. Schulze and Scuttle, (2020) for instance examined the spatial distribution of SOC under natural vegetation and agricultural land use. They concluded that about 50% of the SOC sink in natural vegetation have been lost to intensive cultivation. In another study, Venter *et al.,* (2021) modelled the variability of the topsoil SOC within South Africa's natural vegetation (biomes) and concluded that although the natural vegetation sequesters about 5.6 Pg of organic carbon, a significant amount of the sink is lost to land use change overtime. Despite these studies highlighting the importance of land use on SOC reserves, a deeper understanding of influential land uses and their competing spatial dynamics (e.g. forest plantations and urban vegetation) is required to fully understand the effect of land use change on SOC variability. More importantly, most existing SOC studies are localized with little national, regional and global impact; thus, knowledge on SOC variability due to land use change at the national scale is important for advancing appropriate management strategies (Chaplot *et al.*, 2010). Such knowledge is also critical to achieve among others the total annual national and global carbon accounting objectives, national climate policies, soil management strategies, integrated land use planning and Intergovernmental Panel on Climate Change (IPCC) and Kyoto protocol objectives (IPCC 2016; Zhou *et al.*, 2021; Odebiri *et al.*, 2020a).

The Intergovernmental Panel on Climate Change Good Practice Guidance (IPCC-GPG) on land use change recommends remote sensing (RS) as a highly robust, cost-effective and reliable strategy for mapping different carbon pools, thus, contributing to the long-term climate change regulatory policies (Gara *et al.*, 2016). Whereas conventional field and laboratory methods of SOC determination can provide quality data and information for localised areas, they are impractical at a landscape level. In this regard, RS methods offer the opportunity to provide updated, consistent, and spatially explicit assessments of SOC and its dynamics, especially for large spatial extents with limited access (Mngadi *et al.*, 2019; Xu *et al.*, 2019). This can be credited to recent sensor advancements offering higher spatial and spectral resolutions as well as the multiplication of image datasets (Odindi *et al.*, 2016; Mutanga *et al.*, 2015). However, the performance of RS data as a tool for earth observation depends on many pre-processing procedures. Moreover, because of the heterogeneous nature of RS data (such as high dimensionality), model performance is often dependent on the adopted algorithm (Odebiri *et al.*, 2021). As such, the RS community is constantly searching for innovative analytic techniques to ameliorate the utility and performance of RS data (Masemola and Cho 2019).

At the core of these advancements is the emergence of deep learning (DL) algorithms, which has proven to be an exceptionally powerful tool in many fields (Yuan *et al.*, 2020). DL is part of the broader machine learning (ML) segment based on neural network frameworks (Ma *et* *al.,* 2019). It comprises of neurons, also known as units with many layers that transform input data (e.g. remotely sensed data) to outputs (e.g. estimated SOC), while steadily learning higher-level features (Schmidhuber, 2015; Litjens *et al.,* 2017). Unlike other geostatistical and conventional ML algorithms, DL frameworks can exploit feature representations exclusively learned from data (Odebiri *et al.,* 2021). In addition, DL methods have the capability of improving learning processes, especially with regards to the complex interrelationships between different environmental attributes (Odebiri *et al.,* 2021; Ma *et al.,* 2019). Despite these advantages, DL algorithms have rarely been adopted in SOC retrieval tasks (Singh and Kasana, 2019). Hence, exploring DL applications for remote sensing of SOC could be beneficial to developing more reliable predictive models for frequent carbon assimilation/emission studies. To this end, we predict the spatial distribution of SOC across South Africa's major land uses that include natural forests, commercial forests, urban vegetation, shrub land, grasslands, croplands and barren lands using the recently launched Sentinel-3 satellite imagery and Deep Neural Networks.

5.2. Methodology

5.2.1. Soil data

The soil data used in this study was sourced from both the International Soil Reference Information Centre (ISRIC) and the Department of Agricultural Earth and Environmental Sciences (SAEES) at the University of KwaZulu-Natal in South Africa. ISRIC is an independent science foundation whose mission is to provide worldwide high-quality information on various soil properties (including SOC) through cooperation with global nations. The existing ISRIC soil database, last updated in 2020 (https://www.isric.org/), contains more than 150,000 sample points from 173 countries (Batjes *et al.,* 2020). Most of these sample points differ in space and time of acquisition. In addition, methods for determining SOC content, including field and laboratory spectroscopy, vary between countries (Venter *et al.,* 2021; Hengl *et al.,* 2017). As a result, ISRIC developed a standardized and harmonized procedures for uniform soil profile data input, which are publicly available (https://www.isric.org/explore/wosis/accessing-wosis-derived-datasets). The SOC stock for each point was determined using the following formula by Pearson *et al.,* (2007).

$$SOC\ accumulation\ (t/h) = SOC\ concemtration\ x\ volume\ density\ x\ soil\ depth \qquad (1)$$

A total of 1936 sample points were used as target variable at a 30 cm depth.

5.2.2. Image data acquisition

The study utilized the recently launched Sentinel-3 Ocean and Land Colour Instrument (OLCI) image data, an extension to Sentinel-2 in the Copernicus program by the European Space Agency (ESA). Sentinel-3 incorporates 21 spectral bands within 400nm and 1020nm; with two spectral bands (Oa10 and 11) located in the red edge region (Vaudour *et al.,* 2019). Its 300m spatial resolution, large spatial coverage of 1270 km, and short revisit period (i.e. less than two days) allow for a reliable, timely, and continuous mapping at national, regional, and global scales (Li and Roy, 2017). Furthermore, Sentinel-3 is purposed for evaluating climate change effects and contains spectral bands for bare soil monitoring (Li *et al.,* 2021). The sensor is also important for monitoring vegetation health and conditions due to its strong chlorophyll absorption index (Kokhanovsky *et al.,* 2019). Despite these attributes, Sentienl-3 data is still relatively underexplored within digital soil mapping, hence its choice. Four 300m image tiles with less than 10% cloud coverage captured between March 2021 and April 2021 were downloaded for pre-processing and analysis. The Sentinel Application Platform (SNAP v. 6.0) was used to pre-process the image tiles for geometric, radiometric and atmospheric corrections. Utilizing the Environment for Visualizing Images software (ENVI 5.2), the pre-processed images were mosaicked into one tile, and the country's bounds extracted using ArcGIS 10 (ESRI, 2011). Several studies have demonstrated how vegetation indices are useful in explaining variability and distribution of SOC (Wang *et al.,* 2021; Odebiri *et al.,* 2020b; Guo *et al.,* 2020; Dotto *et al.,* 2018). Hence, ten relevant vegetation indices were generated using different combinations of Sentinel-3's twenty-one spectral bands. The specification and justification for these indices are summarized in Table 5.1. The final SOC model was generated using both the 21 Sentinel-3 spectral bands and the ten vegetation indices developed from the sensor.

Table 5.1. Sentinel-3-derived spectral vegetation indexes

Indices	Justification for use	Formula	Reference
Normalized Difference Vegetation Index(NDVI)	Calculates the vegetation density by separating soil from vegetation, and minimizes topographical variation	$\frac{NIR - RED}{NIR + RED}$	Rouse (1974)
Soil Adjusted Vegetation Index (SAVI)	Reverses the effects of low vegetation	$\frac{NIR - RED}{NIR + RED + 0.5}(1 + 0.5)$	Huete (1988)

	cover on soil brightness		
Optimized Soil Adjusted Vegetation Index (OSAVI)	Adaptable to greater soil changes	$(1 + 0.16)\ (NIR - RED)/(NIR + RED + 0.16)$	Jamalabad and Abkar , (2004)
Modified Soil Adjusted Vegetation Index (MSAVI)	Addresses areas with high soil surface exposure	$\frac{2NIR + 1 - \sqrt{(2NIR + 1)^2 - 8(NIR - RED)}}{2}$	Qi (1994)
Enhanced Vegetation Index (EVI)	Sensitivity and signal enhancement in vegetation-rich areas	$2.5 \times \frac{NIR - RED}{(\mathrm{NIR} + 6 \times \mathrm{RED} - 7.5\ \times \mathrm{BLUE} + 1)}$	Huete (1999)
Ratio Vegetation Index (RVI)	Vegetation amount is indicated	$\frac{NIR}{RED}$	Baret (1991)
Renormalized Difference Vegetation Index (RDVI)	Connects the biophysical parameters to the index in a linear way	$\frac{(NIR - RED)}{(NIR + RED)^1/2}$	Roujean and Breon (1995)
Transformed Vegetation Index (TVI)	Sensitive to the amount of vegetation	$\sqrt{(NDVI)} + 0.5$	Deering (1975)
Difference Vegetation Index (DVI)	Differentiates between soil and vegetation	$NIR - RED$	Richardson (1977)
Green Normalized Difference Vegetation Index (GNDVI)	Sensitive to vegetation and chlorophyll content in plants	$\frac{NIR - GREEN}{NIR + GREEN}$	Gitelson (1998)

5.2.3. The SOC Model

A deep neural network (DNN) workflow was utilized in modelling the spatial distribution of SOC across important land uses in South Africa. DNN is a powerful forward-feeding machine learning mechanism that provides information about linear and non-linear connections between the response variable and a set of predictor variables (Wang *et al.,* 2020). Intrinsically, DNN uses a multilayer perceptron architecture (MLP), but its hidden layers and built-in hyper-parameters make it superior to other geostatistical and classical machine learning models (Taghizadeh-Mehrjardi *et al.,* 2020). The input, the hidden layer, and the output layer comprise of several neurons. The neurons in each layer are connected to every neuron within each subsequent layer until the final output (i.e. SOC). In recent years, DNN has gained attention in the field of remote sensing, including SOC mapping (e.g. Emadi *et al.,* 2020; Taghizadeh-Mehrjardi *et al.,* 2020). For a given input layer (vector X) with L hidden layers and an output layer (vector Y), the mathematical representation of DNN is as follows (Wang *et al.,* 2020);

$$Z^1 = \sigma_1(W^1X + b^1), Z^2 = \sigma_2(W^2z^1 + b^2)$$

$$Z^L = \sigma_L(w^Lz^{L-1} + b^L), \;\; Y = \; W^{l+1}Z^L + b^L, \theta = [W^i, b^i]_{i=1}^{L+1} \tag{2}$$

The weight and bias of the ith layer are represented by W^i and b^i, respectively. $L + 1$ indicate the output layer (i.e. Y = N (X; Θ)) and σ^i is the activation function of the ith layer which can either be sigmoid, rectified linear unit (ReLU) or hyperbolic tangent (Tanh).

By using the mean square error (MSE), we can examine the loss function *L* of the input and output variables as expressed below;

$$\mathrm{MSE_{DATA}} = \mathrm{L}(\theta) = \frac{1}{\mathrm{N}}\sum_{i=1}^{N} |\mathrm{NN}(X_i;\; \theta) - Y_i|^2 \tag{3}$$

In this case, *N* denotes the number of the labelled data. The loss function *L* can be minimized through an optimization algorithm called Stochastic gradient descent (Wang *et al*., 2020). Figure 5.1 illustrates the DNN architecture. A total of 31 input variables, including 21 spectral bands and 10 vegetation indices from Sentinel-3 OLCI, were used for the final SOC model. The dataset (*n* = 1936) was divided into training (*n* = 1084), validation (*n* = 271), and testing (*n* = 581) data using a 50%, 20% and 30% split. To avoid overfitting, the model during training, a regularization technique using the dropout function was applied to each of the layers, which were then averaged for prediction. Consequently, following repeated adjustments carried out by the random search technique adopted for hyper-parameter optimization, six hidden layers, the "Adam" optimizer and the ReLU activation function were selected for optimal results.

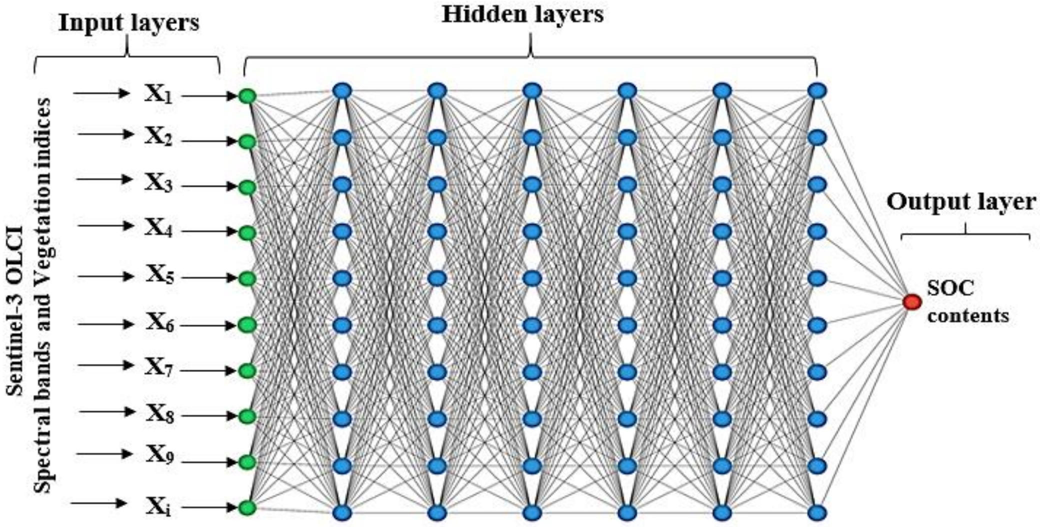


Figure 5.1. A graphical illustration of the deep neural network (DNN) framework used in the study

5.2.4. Model evaluation and uncertainty

This study evaluated the fit and generalization of the SOC DNN model through three different metrics. These included the root mean square error (RMSE), the coefficient of determination (R^2) and Lin's concordance correlation coefficient (LCCC). These evaluation metrics are expressed as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_{O,i} - X_{P,i}\right)^2}{n}} \tag{4}$$

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} (X_O - p)^2}{\sum_{i=1}^{n} (X_O - O')^2}\right] \tag{5}$$

$$\mathrm{LCCC} = \frac{2\, r\, \sigma_O \sigma_p}{\sigma_p^2 + \sigma_p^2 + [O' - P']^2} \tag{6}$$

where, *n* connotes the number of observations, X_O and X_P are the measured and predicted SOC value, respectively. *O'* and *P'* reflect the averages of the measured and predicted SOC, and $\boldsymbol{\sigma}_o$ and $\boldsymbol{\sigma}_p$ reflect the variances of observed and predicted values. Furthermore, a tenfold cross-validation approach was applied to ensure that the model did not over-fit the data (Mutanga *et al.*, 2012). The SOC data was subsequently divided into a set of 10, with calibration and validation data successively added until each variable had been utilized. In general, the most suitable model has a higher R^2 and LCCC value and a lower RMSE value. Moreover, the importance of each predictor to the overall performance of the DNN model was assessed. Although deep learning (DL) models often cannot automatically achieve interpretability, several approaches (such as the SHapely Additive exPlanations method) have been recently proposed to help users interpret DL models (Pentoś 2016). Hence, in this study, the SHapely Additive exPlanations (SHAP) technique was used to evaluate predictor variable importance. SHAP works by assigning a specific mean value to each predictor variable, which then indicates the magnitude of their effect on the model output. SHAP has special features such as DeepExplainer (for DL-based methods), TreeExplainer (for tree-based methods), and KernelExplainer (for other model types) that correspond to all types of ML models. Furthermore, SHAP offers both global and local interpretability (see Lundberg and Lee, 2016 for a more detailed explanation of SHAP).

Along with evaluating a model's performance, it is also important to quantify its uncertainty (Abdar *et al.*, 2021). According to Hamel and Bryant, (2017), a study that is decision-oriented must identify and evaluate the uncertainty and robustness of the study conclusions to make the analysis credible. Thus, to quantify the level of uncertainty within this study, the upper and lower bounds were provided for the SOC model (Abdar *et al.*, 2021). This was determined by utilizing the ±1.64 standard deviation (SD) with a 90% confidence interval (C1) and a 10-fold cross-validation (Emadi *et al.,* 2020; Minasny *et al.,* 2016). The 5th and 95th percentiles were then retrieved along with the predicted mean value of each pixel. As a final step, a spatial distribution map was generated for the calculated mean, lower (5%) and higher (95%) confidence intervals.

5.2.5. Land use class determination

Following the SOC model's generation, the spatial variability of SOC across seven different land use classes within South Africa were examined. These classes include natural forest, commercial forest, grassland, cropland, urban vegetation, shrub land and barren land. These land uses were specifically chosen due to their dynamic influence over SOC stocks. The land cover types and their spatial distribution were derived from the South Africa National Land-Cover Map (SANLC, 2020). The map, with an 80% accuracy, was developed using 20m Sentinel-2 satellite imagery over a multi-temporal time period. Full details of the land use determination procedure and the data used is freely available through the E-GIS website (https://egis.environment.gov.za/gis_data_downloads). The seven land use classes used in this study were individually extracted from the SANLC image using the ArcGIS pro 2.7 API. Thereafter, the extent of each land use within South Africa was extracted from the DNN SOC map. The amount of SOC and area occupied by each land use class was derived through the calculate geometry tool within the ArcGIS environment. In addition, descriptive statistics (minimum, maximum, mean and the total SOC stocks) for each class was attained. Figure 5.2 and Table 5.2 illustrate the spatial distribution of the different land use classes as well as their respective descriptions.

Table 5.2. Land use classes and their description derived from the SANLC, 2020

Land use type	Description
Natural forest	Naturally growing forest including contiguous (indigenous) forest, contiguous low forest and thicket, dense forest and woodland, open woodland.
Commercial forest	Planted forest for economic purposes including contiguous and dense plantation forest, open and sparse plantation forest, temporary unplanted (clear-felled) plantation forest
Grassland	Natural grass land and sparsely wooded grassland
Urban vegetation	Urban recreational fields including trees, bush, grass and bare in built-up areas
Cropland	Cultivated permanent crops, temporary crops and fallow lands/old fields
Barren land	Natural rock surfaces, dry pans, eroded lands, sand dunes (terrestrial), coastal sand and dunes, bare riverbed material, and other bare surfaces
Shrub land	Natural low woody communities with a canopy height ranging between 0.2 to 2 meters

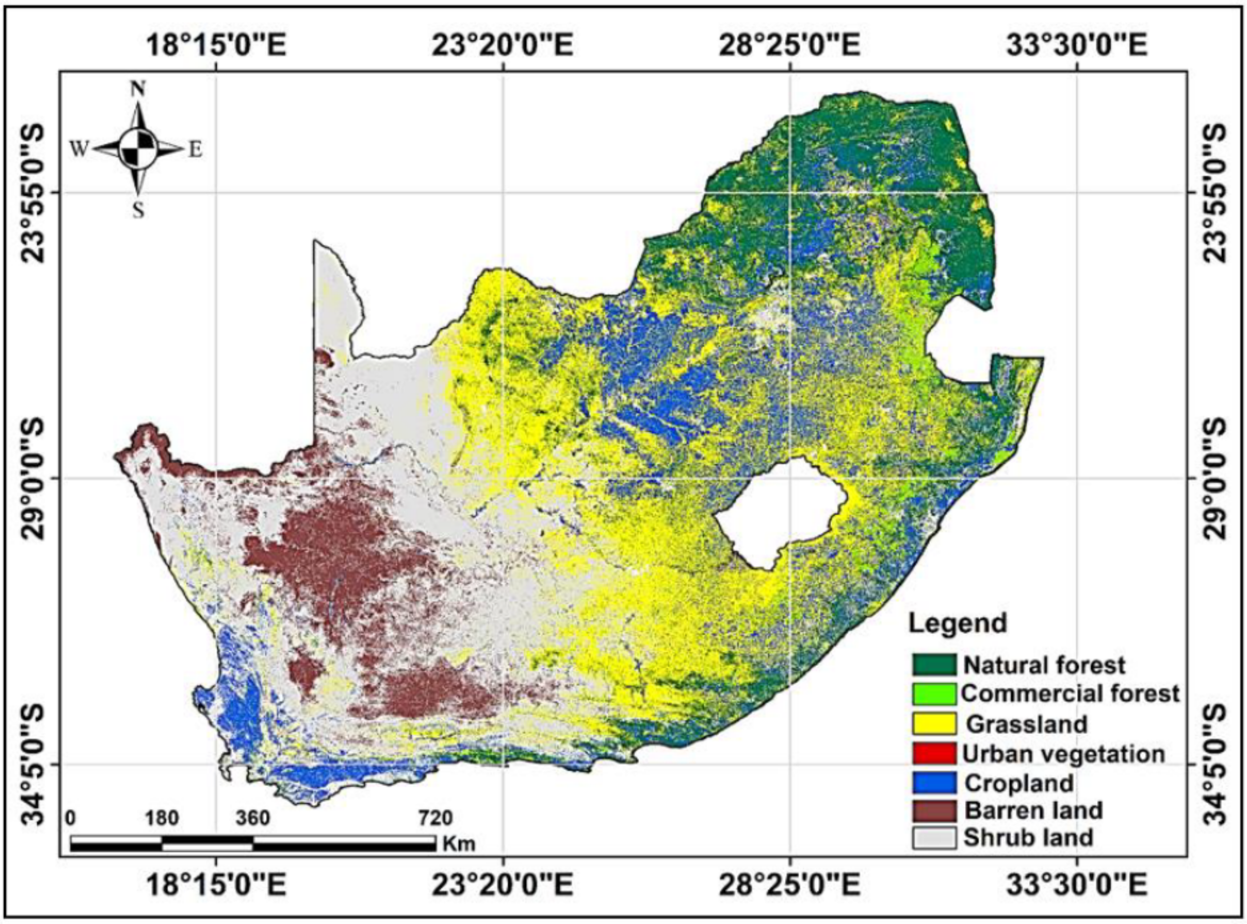


Figure 5.2. Land use classification map showing the seven different land uses; including natural forest, commercial forest, grassland, urban vegetation, cropland, shrub land and barren land (SANLC, 2020)

5.3. Results

5.3.1. Descriptive statistics

The SOC data (1936 soil samples) ranged from 5.3 t/h to 149 t/h, with a mean of 39.8 t/h and a standard deviation of 17.3 t/h. The 43% SOC variance, defined as the ratio of the standard deviation to the average, was relatively high. In addition, the data demonstrated high skewness (1.9) and kurtosis (5.2) as it deviated from the normal distribution curve (Hair *et al.,* 2016). To improve the distribution of the data, we applied a natural logarithm transformation technique, which generated new skewness and kurtosis values of 0.41 and 0.68, respectively. A reverse transformation of the transformed SOC data was performed to restore the data to its original scale after predictive analysis.

5.3.2. Evaluation and performance of the DNN model

The SOC DNN model, developed from 21 spectral bands and 10 vegetation indices, performed relatively well in estimating SOC stocks across South Africa, with an R^2 value of 0.685 and an RMSE of 10.15t/h (Figure 5.3B). Moreover, both train and test data demonstrated a good model fit and generalization (Table 5.3). Specifically, training data generated an R^2 value of 0.685, an RMSE of 10.15t/h (26% of the mean) and an LCCC value of 85.46 (Table 5.3). Meanwhile, testing data revealed R^2, RMSE and LCCC values of 0.642, 11.96 t/h, and 79.67, respectively (Table 5.3). Figure 5.3A shows the loss curve for train and validation data, indicating that there was no overfitting in the model. Furthermore, the SHAP technique was used to determine the significance of each predictor variable to the DNN model. The five most important explanatory variables were Band 8 (665 nm) located in the red spectrum region, NDVI, Band 11 (708.25 nm) located in the red-edge spectrum region, EVI and RVI (Figure 5.3C). Specifically, Band 8 and NDVI had a similar magnitude of influence over the SOC model (Figure 5.3C).

Table 5.3. The SOC DNN model result summary for both train and test data

Evaluation metrics	Train data	Test data
R^2	0.685	0.642
RMSE (t/h)	10. 15	11.96
LCCC	85.46	79.67

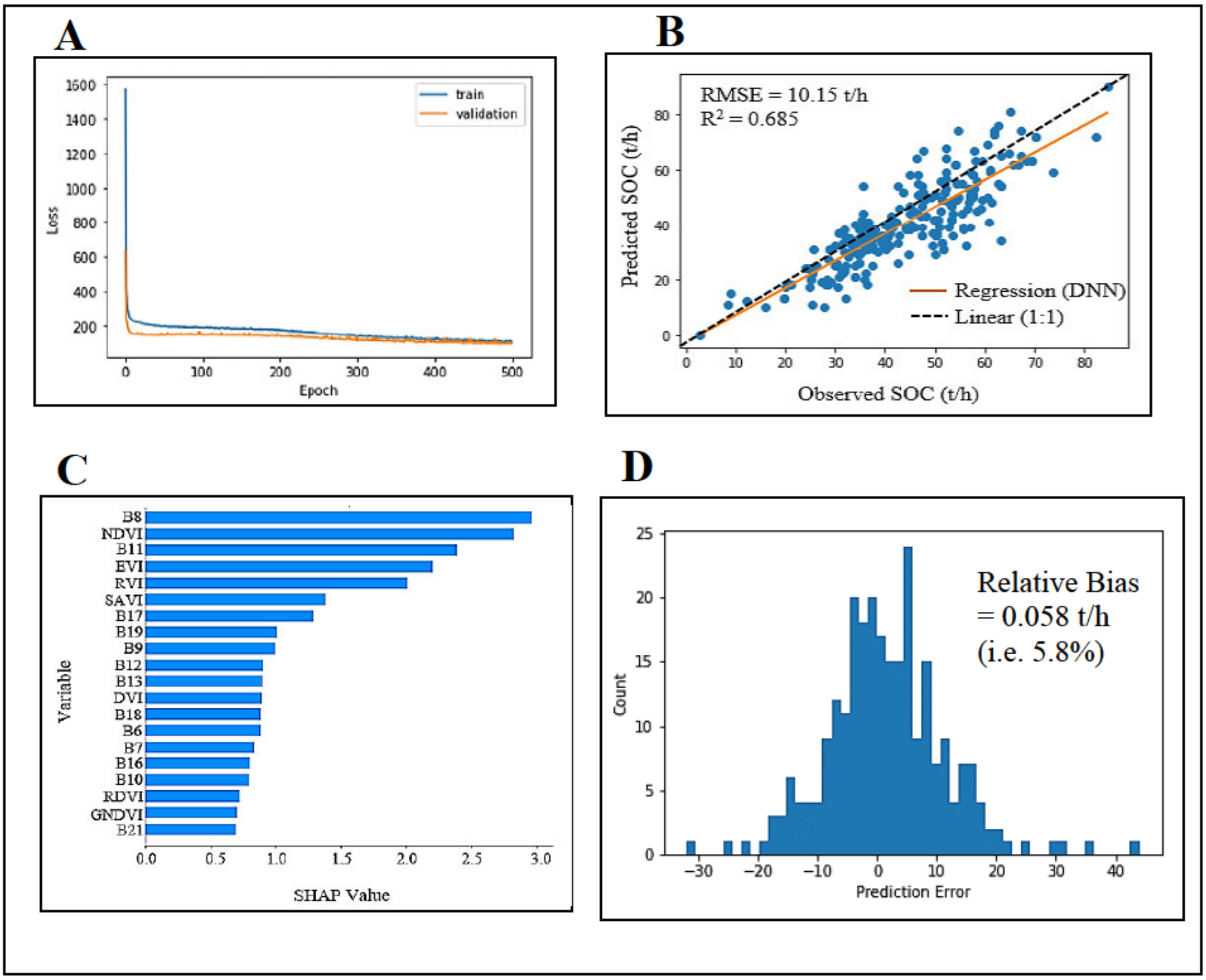


Figure 5.3. (A) shows the training and validation loss, (B) correlation between the predicted and observed soil organic carbon (SOC) using the training data, (C) rank of predictors used to simulate SOC across land use types; the rankings were generated using the SHAP technique, (D) a bias plot showing a balanced distribution and a relatively unbiased model

5.3.3. Spatial estimation of SOC and uncertainty quantification

Figure 5.4 illustrates the spatial distribution and uncertainty quantification of predicted SOC for South Africa at the upper limit (95%), the mean, and the lower limit (5%), calculated at a 90% confidence interval. All the maps show consistency. From the mean map, concentrations of SOC are generally higher in densely vegetated areas in comparison to those with sparse or no vegetation. For instance, the north-east and south-east parts with darker colours are dominated by dense natural forests and plantations, whereas the arid western areas are dominated by wastelands. In addition, our DNN model provided a reliable uncertainty estimate with 88% of the observations falling within the defined confidence interval. However, it is important to emphasize that although the DNN model depicts a good measure of uncertainty, it may not reflect the actual situation on the ground, since the quantification of uncertainty here relies on hyper-parameters of the model rather than the spatial uncertainty of the data.

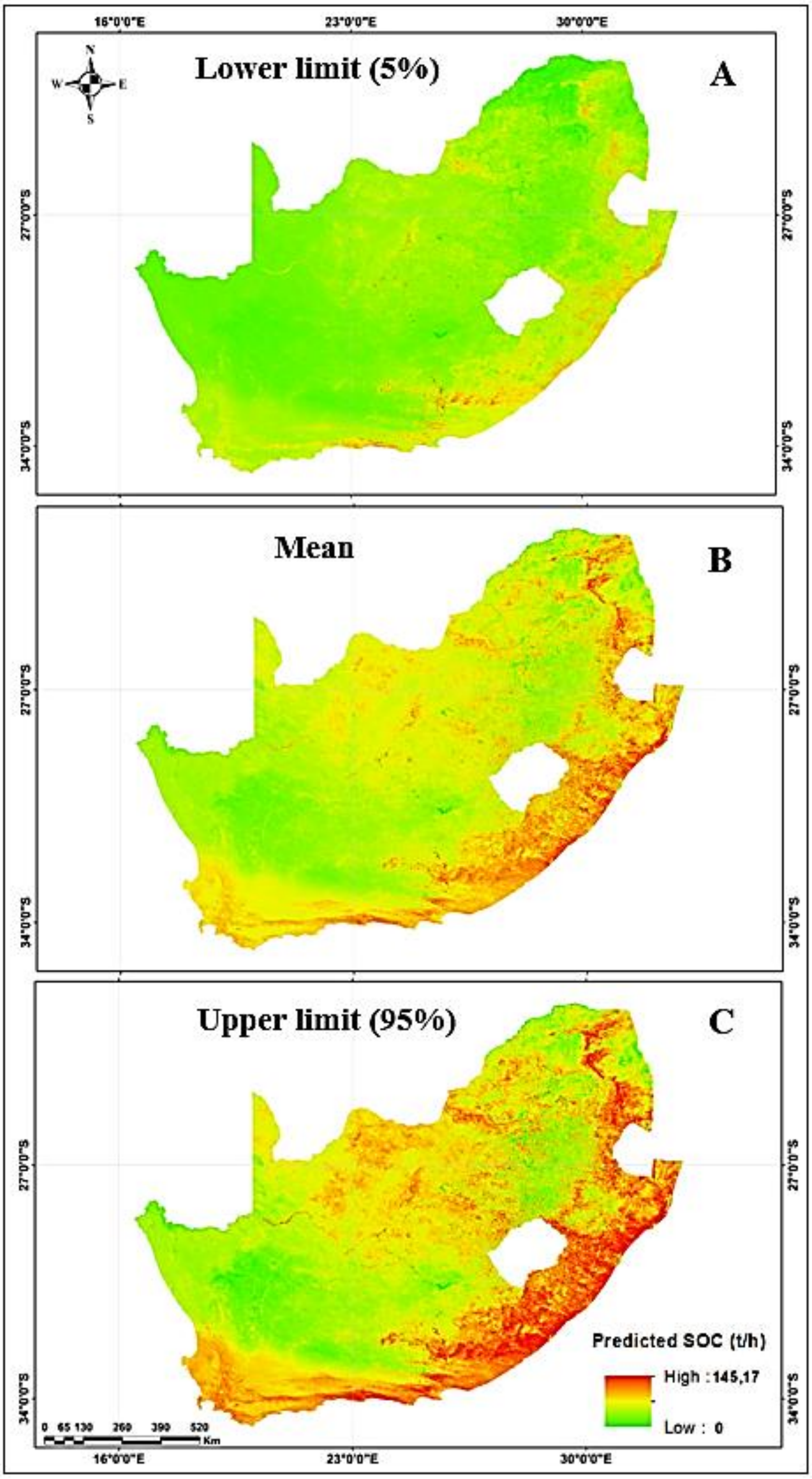


Figure 5.4. Spatial estimation of soil organic carbon at (A) lower limit (5%), (B) mean, (C) upper limit (95%) derived using Sentinel 3 OLCI data and deep neural networks (DNN)

5.3.4. SOC assessment across different land use types

The summary statistics as well as the spatial distribution of SOC across the different land use types are depicted in Table 5.4 and Figure 5.5. The total area of the seven classes is approximately 1,188,033 km^2 with the grassland occupying roughly 29%, shrub land 27.33%, natural forest 15.87%, cropland 15.04%, barren land 11.15%, commercial forest 1.74% and urban vegetation 0.04% (Table 5.4). Grasslands have the highest total concentration of SOC stocks across the different land use types with 31.36%, while urban vegetation has the least with 0.04% of the total SOC stocks in South Africa. Natural forest has the highest maximum concentration in relation to a single soil sample point with 145.17 t/h, while barren land has the least with 6.3 t/h. In addition, the mean values of SOC stocks for each land use type provides an indication of the carbon sequestration potential of each land use in respect to their total land area. Specifically, commercial and natural forests, have the highest average SOC stocks concentrations of 46.06 t/h and 44.34 t/h, respectively. This is significantly greater than the larger grassland class — with an average SOC stock of 31.72 t/h. Urban vegetation, despite its small area, has a relatively high average SOC stocks concentration (33.72 t/h) when compared to cropland (33.46 t/h), grassland (31.72 t/h) shrub land (22.6 t/h). Meanwhile, barren land showed the lowest SOC stock retention potential, with 20.41 t/h. It is spatially evident that the majority of South Africa's SOC stocks are mostly distributed along the east coast of the country (Figure 5.4B). Moreover, higher SOC concentrations for each land use (apart from barren land and shrub land) are mostly prevalent within KwaZulu-Natal, Eastern Cape, and Mpumalanga provinces (Figure 5.5). SOC Stocks for barren land and shrub land are mostly found within the Western and Northern cape provinces.

Table 5.4. Summary statistics of the spatial distribution of SOC stocks across important land use types in South Africa. Values in square parentheses are the 5th and 95th percentile values

Land class type	Class area (%)	Total SOC (%)	Min SOC (t/h)	Mean SOC (t/h)	Max SOC (t/h)
Grassland	28.83	31.36	9.096 [4.8, 13]	31.72 [27.6, 36]	87.44 [85, 94]
Shrub land	27.33	20.97	6.5 [2.7, 11]	22.6 [18.7, 27.1]	60.6 [57, 65]
Natural forest	15.87	19.41	12.56 [8, 16.4]	44.34 [38.6, 48.3]	145.17 [139, 149]
Cropland	15.04	17.02	6.87 [2.8, 11]	33.46 [29, 37.4]	86.88 [84, 93]
Barren land	11.15	8.45	6.30 [2.3, 10.3]	20.41 [16.4, 24.4]	58.95 [54, 61.4]
Commercial forest	1.74	2.75	14.61 [10.5, 19]	46.06 [42, 50.8]	98.55 [94, 103]
Urban vegetation	0.04	0.04	28.31 [24.3, 32.2]	33.72 [29.8, 38]	44.72 [41, 50]

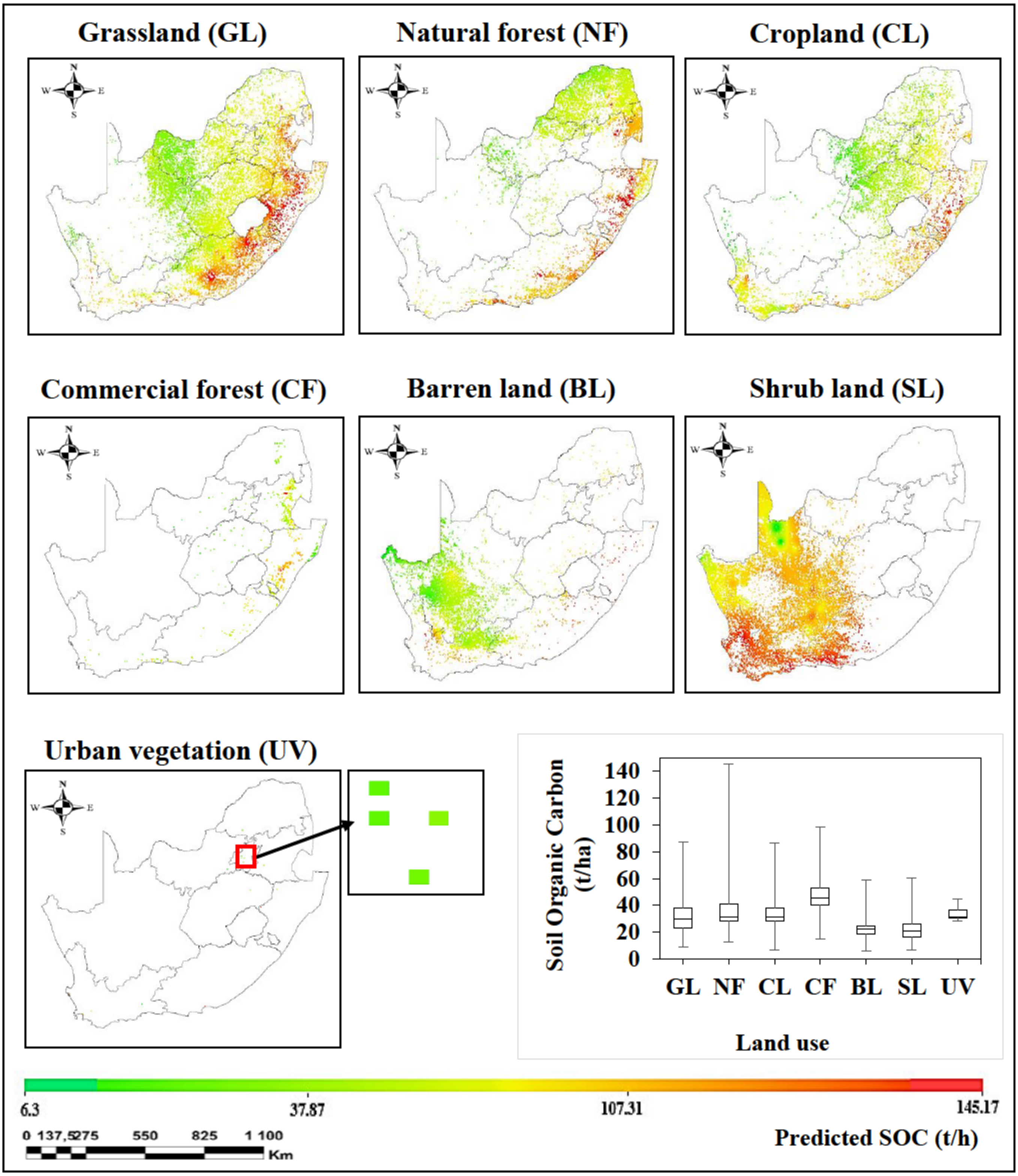


Figure 5.5. Spatial distribution of SOC stocks across South Africa's important land uses; The urban vegetation class is invisible at full extent; hence, a zoomed-in section in the red box of the class is depicted within the adjacent box pointed by the black arrow. The box plot within the map also depicts the spatial variability of SOC between the land use classes

5.4. Discussion

Soil organic carbon (SOC) is a key indicator of soil and vegetation health and crop efficiency (Wang *et al.,* 2021). Its large reserves in terrestrial ecosystems provide important opportunities to mitigate climate change and regulate carbon flux (Odebiri *et al.,* 2020b). However, SOC is highly influenced by vegetation through organic and inorganic inputs, hence land use change is one of the most important determinants of its accumulation (Kenye *et al.,* 2019). Whereas, a land use type and change that exerts the least soil disturbance may contribute to increased SOC accumulation, intensive disturbance results in SOC loss and lower uptake (Ghimire *et al.,* 2018). Consequently, understanding and modelling the spatial distribution of SOC across different land use types is critical to advancing appropriate management practices to improve soil quality and carbon sequestration potential in mitigating climate change and associated impacts.

5.4.1. Spatial distribution of SOC across different land use types

Results in this study show that grasslands have the highest concentration of SOC stocks (31.36%). This is in agreement with Venter *et al.*'s (2021) findings that South Africa's grassland biomes have higher stocks of SOC (36%) than other landscapes. The high grasslands SOC reserves are due to their large areal extent (28.13%). Furthermore, South Africa's grasslands are mostly located within the central, northern, and eastern parts of the country (Figure 5.5), characterized by higher precipitation (>500 mm). Rainfall influences soil moisture, hydrological processes (such as surface runoff and ground water infiltration), vegetation density, and decomposition, which contribute to SOC sequestration (Zhou *et al.,* 2008; O'Brien *et al.,* 2010; Chen *et al.,* 2015). Globally, grasslands store between 12 and 18 percent of all terrestrial carbon stocks (Conant *et al.,* 2001, Ontl, 2017). Unlike other ecosystems, grasslands store most of their sequestered carbon underground within the rooting zone, making them more resilient to natural environmental disturbances such as wildfires (Ward *et al.,* 2016). In addition, certain grassland ecosystems (such as the KwaZulu-Natal Sandstone Sourveld) contain forbs and *Geoxylic suffrutices* that retain greater quantities of carbon within their woody biomass, thus contributing to the overall carbon stocks (Gomes *et al.,* 2021; Zaloumis and Bond, 2016; Ampleman *et al.,* 2014) (Table 5.2). There is, however, some agricultural disturbance to South Africa's grasslands since they are endemic to arable nutrient-rich land, which makes them ideal for crop cultivation (Little *et al.,* 2015; Palmer and Ainslie, 2005). Consequently, 60 % of South Africa's grasslands have been irretrievably transformed, with only 2 % under official conservation (Little *et al.,* 2015). Given these competing land uses, adequate spatial planning and management frameworks directly supported by scientific data is required to secure SOC stocks within South African grasslands. An example of such effort is the Durban Research Action Partnership (D`RAP), which is a national science scheme that combines the eThekwini Municipality's environmental management department with researchers from the University of KwaZulu-Natal to scientifically inform and supplement applied environmental monitoring and management protocols (Palmer and Ainslie, 2005; Cockburn *et al.,* 2016). D`RAP supported scientific research has enabled critically endangered grassland patches and other ecosystems to be included within the Durban Metropolitan Open Space System (DMOSS) framework, which is an instrument used to ensure that biodiversity concerns are integrated into national development planning (Roberts *et al*., 2012). Moreover, studies such as Rouget *et al.,* (2016), have recommended that critically endangered grassland patches be included within the South Africa climate change strategy.

A comparison of the minimum, average, and maximum SOC stock values (Table 5.4) reveals how storage rates vary across each land use type. For instance, although barren and shrub land cover a large area, they have the lowest minimum, average, and maximum values compared to other land uses. These low SOC values are related to relative aridity and low vegetation cover (which reduces carbon sequestration capacity), as most of the barren and shrub land is distributed within the semi-arid, arid and deserted regions of South Africa (Venter *et al.*, 2021). It is important to note that shrubland accounts for a significant percentage (20.97%) of the total SOC reserves due to the large area covered, second to the grassland biome. Barren land also had a considerable total SOC (8.45%) and maximum value (58.95 t/h), which was unexpected. This may be related in part to the large area coverage (11.15%) or the misclassification of the SANLC (2020) data used in this study since the classification accuracy is not 100%. Forest ecosystem including natural and commercial forests had the highest mean (46.06 t/h), and maximum (145.17 t/h) SOC values, indicating their ability to sequester more carbon than other land uses, despite their limited spatial extent (Table 5.4). Forested areas are characterized by tall indigenous and exotic trees with longer rooting residency, canopy coverage and height ranging from 10-75% and 2.5-6 meters, respectively (SANLC, 2020). This favours the rate of SOC sequestration due to accelerated soil metabolism from continuous litter fall and dead matter (Muchena, 2017). Additionally, commercial plantations are often extensively managed to balance economic (e.g. wood supply) and environmental needs (Odebiri *et al.,* 2020b).

Natural forests on the other hand, sustain biodiversity and provide essential ecosystem services (Department of Environmental Affairs, 2017). However, the need for housing and agricultural land has driven deforestation across vast swaths of land, leading to SOC stocks loss (Leblois *et al.,* 2017). While this calls for indigenous reforestation projects which are vital to sustaining ecosystem services and climate change mitigation (Abiodun *et al.,* 2012), care must also be taken to balance trade-offs between carbon sequestration, biodiversity and the overall ecosystem function (Bond *et al.,* 2019; Venter *et al.,* 2021).

Cropland accounts for 15% of total SOC stocks, with a mean value of 33.46 t/h, which indicates a relatively high sequestration rate. Croplands comprise of predominantly permanent and temporary crops, including pineapples, sugarcane, and vines (SANLC, 2020). The crops remain in-field for multiple growing and harvesting seasons, and are generally located in areas of higher rainfall (e.g. northeast and southeast zones), which facilitates the accumulation of SOC. In areas with less rainfall, such as those in the western part of the country, agricultural crops may benefit from irrigation and fertilizer application, thereby increasing the potential for the sequestration of SOC. Literature shows that addition of soil nutrients contributes to SOC sequestration in croplands by increasing biomass production and the carbon-nitrogen ratio (C: N) (Tiefenbacher *et al.,* 2021 Hijbeek *et al.,* 2019; Han *et al.,* 2016). In addition, cropland areas also include fallow fields which were previously cultivated and are now overgrown with trees, bushes, grasses and shrubs, thereby increasing the SOC pool and the rate of sequestration (Yeasmin *et al,* 2020, Sharma *et al.,* 2019). Howbeit, it is vital to note that continuous cultivation has been noted to significantly reduce South Africa’s SOC reserves (Venter *et al.,* 2021; Schulze and Scuttle, 2020; Swanepoel *et al.,* 2016). Besides, the prevalence of subsistence farming especially in rural areas, where agro-forestry, crop rotation and fallow systems are rarely practiced, continues to contribute significantly to SOC loss (Rusere *et al.,* 2019; Khapayi *et al.,* 2016). Although cultivation is inevitable to provide food security for the increasing population, it is necessary to balance productivity with overall ecosystem functions; including climate change mitigation. Comprehensive educational-training programmes that promotes sustainable farming system must also vigorously engage rural areas to foster a holistic land use management and monitoring scheme.

The urban vegetation class as specified by SANLC (2020), comprises of trees, grasses, bushes, and shrubs. Despite being the smallest land use class, the urban vegetation shows a relatively high rate of SOC sequestration (mean = 33.72 t/h). This suggests that urban vegetation could be critical to offsetting urban carbon footprints and thus reduce the urban heat island effect in cities. Ideally, urban green and open spaces should form a major component within integrated urban planning (Beecham *et al.,* 2019). Examples of this can be found within the tree-planting initiative in Pretoria (Tshwane) — where 55 000 tons of carbon are sequestered each year (Grebner *et al.,* 2012), and the Durban Research Action Partnership (DRAP) Community reforestation project in eThekwini, which has an average carbon sequestration rate of 2110.7 $tCO_2e.yr^{-1}$ for the next 20 years (CSIR 2006; Ethekwini municipality and Wildland's conservation trust, 2014).

In general, although interventions to restore degraded ecological areas are to be commended and encouraged, we caution against efforts to transform other naturally existing land uses into forests as motivated by the global carbon market (Venter *et al.,* 2021; Moyo *et al.,* 2021; Bond *et al.,* 2019; Mills and Cowing 2014). Altering the natural state of vital ecosystems, such as grasslands, will be ecologically devastating to biodiversity and may be counterproductive. Moreover, SOC stocks previously sequestered within forests are vulnerable to loss from forest fires. Therefore, in the wake of intensified land-use conflicts, the demands of a growing population, and the existential threat of climate change, evidence-based research in partnership with sound land-use planning is required to provide an appropriate balance between existing land uses. In addition, innovations in modern satellite sensor technology and expedited research into the development and use of artificial intelligence may provide the tools required for the better understanding of spatial SOC dynamics.

5.4.2. Performance of the SOC DNN model

This study developed a spatial SOC variability map based on Sentinel-3 satellite data and deep learning approach (DNN). It is the first national SOC model developed using deep learning framework in South Africa ($R^2 = 0.685$). Considering the spatial uncertainty of the reference data (i.e. multiple-scales of variation, various sampling sources and different collection times), our DNN model still produced a fairly good result — indicating its ability to model complex data. The model also serves as an improvement to the national scale models previously developed by Venter *et al.* (2021) ($R^2 = 0.659$) and Schultze & Schutte (2020) ($R^2 = 0.203$) using random forest algorithm and field-level SOC median calculation strategy, respectively. The performance of DNN model can be attributed to its ability to learn and extract more representative features from the SOC data through its many hidden layers and neurons (Ma *et al.,* 2019). DNN can accurately approximate the complicated non-linear relationship between SOC and covariates, thus capturing the potential association between them (Yuan *et al.,* 2020). In addition to the robust DNN framework performance, the Sentinel-3 data used in this study was also a good fit for nationwide SOC mapping. This study adopted the use of only spectral information as opposed to the general combination of ancillary data (i.e. topography and climate). This is because we were interested in the performance and influence of RS imagery, specifically Sentinel-3, which has been underexplored despite its impressive spectral attributes and suitability for largescale mappings; and to the best of our knowledge, has never been used as a standalone input data for SOC mapping like other sensors (Bhunia *et al.,* 2017; Guo *et al.,* 2020; Wang *et al.,* 2021). This is motivated by the fact that remotely sensed metrics e.g. vegetation indices can be reliably used to determine vegetation cover, which in turn determines the amount of soil carbon. Generally, vegetation density and distribution is hugely influenced by topographic and climatic characteristics, hence in absence of these datasets, vegetation cover can be used as a topo-climatic surrogate. In addition, Sentinel-3 acquires data in the visible (VIS) to near infrared (NIR) wavelength region of the electromagnetic spectrum (400 to 1020 nm), which is deemed the most sensitive for detecting SOC (Lin *et al.,* 2020; Bilgili *et al.,* 2010). The Sentinel-3 image contains spectral bands for bare soil monitoring, and covers a large area, which is important for national and global scale mapping (Li *et al.*, 2021). It would however be interesting to see how well our deep learning model performs with images of higher spatial resolution like Sentinel-2 and Landsat 8.

As part of a variable importance measure, the most significant variables from the Sentinel-3 derived metrics that explain the distribution of SOC in the study area were evaluated using the SHAP technique (see section 2.4). Band 8 (B8 = 665nm), NDVI, Band 11 (B11 = 708.25nm), EVI and RVI were identified as the most influential variables. Band 8 and 11, which represent the red and red-edge region of the electromagnetic spectrum, are extremely sensitive to vegetative attributes like biomass and chlorophyll content (Mngadi *et al.,* 2019; Forkuor *et al.,* 2017). Subsequently, these bands proved important in SOC mapping due to the correlation between vegetation and SOC concentration (Taghizadeh-Mehrjardi *et al.,* 2020; Zhang *et al*., 2019; Nabiollahi *et al*., 2018). Similarly, NDVI, EVI and RVI enhance plant signals and sensitivity in high biomass areas, supporting the hypothesis that SOC responds to the same physical and environmental factors as biomass (Bhunia *et al.,* 2017; Kumar *et al.,* 2016; Matsushita *et al.,* 2007).

5.5. Conclusion

Using remote sensing and deep learning, this study examined the spatial distribution of SOC across important land use types in South Africa. The results show the estimated SOC stocks within each land use and their potential sequestration rates in relation to area coverage. This is critical to inform national policies, rehabilitation, restoration and intervention efforts in attempts to mitigate climate change and improve soil quality. Despite the spatial uncertainty of the data used in this study, our SOC model based on sentinel-3 data and DNN performed well compared to other national SOC models in South Africa. This calls for standardized soil inventory schemes funded by the government and research agencies where quality data unique to different land use types are collected and made available to improve modelling efforts. Future investigations can consider evaluating SOC stocks distribution and the impact of anthropogenic land use change across distinctive South Africa biomes and bioregions as an alternative to inform management and climate mitigation policies. Further studies can also benefit from the addition of other important SOC environmental drivers including topography and climate to improve accuracy and better understand SOC variability in the area. The DNN model can also be improved in future studies. For instance, this study did not consider a variable selection strategy to reduce possible multicollinearity between predictors which could potentially reduce accuracy.

5.6. Summary

This chapter mapped SOC stocks across South Africa's major land uses using DNN and Sentinel-3 satellite data to examine the impact of different landscapes on SOC stock distribution. Among the different land uses evaluated, grasslands were found to contribute the most to overall SOC stocks, while urban vegetation contributed the least. Although SOC stock was discovered to be relatively proportional to overall land coverage, commercial and natural forests demonstrated a greater carbon sequestration capacity. These findings provide an important guideline for managing SOC stocks in South Africa and will assist in climate change mitigation. However, to adequately comprehend the overall variability of SOC within South Africa, the impact of environmental variables (such as topography and climate) on SOC needs to be understood. Consequently, given South Africa's distinct climatic envelopes and unique biodiversity, the next chapter sought to evaluate the impact of topo-climatic variables and anthropogenic land use change across South Africa's major bio regions as an alternative way to spearhead SOC policies and management. The outcomes of this investigation will provide a *model to contextualize the influence of large bioclimatic zones on regional SOC stock management.*

Chapter Six:
Mapping soil organic carbon distribution across South Africa's major biomes using remote sensing-topo-climatic metrics and Concrete Autoencoder-Deep neural networks

This chapter is based on;

Odebiri, O., Mutanga, O., Odindi, J., & Naicker, R. (2022). Mapping soil organic carbon distribution across South Africa's major biomes using remote sensing-topo-climatic metrics and Concrete Autoencoder-Deep neural networks. *Science of the Total Environment*, Under Review, Manuscript ID: **STOTEN-D-22-10001**

Abstract:

The management of soil organic carbon (SOC) stocks remains at the forefront of greenhouse gas mitigation. However, unprecedented anthropogenic disturbances emanating from continued land-use change have significantly altered SOC distribution across global biomes leading to considerable carbon losses. Consequently, understanding the spatial distribution of SOC across different biomes, particularly at larger scales, is critical for climate change policy formulation and planning. Advancements in remote sensing, availability of big data, and deep learning architecture offer great potential in large-scale SOC mapping. In this regard, this study mapped SOC distribution across South Africa's major biomes using remotely sensed-topo-climatic data and Concrete Autoencoder-Deep Neural Networks (CAE-DNN). From the different deep neural frameworks tested, the CAE-DNN model (developed from 26 selected covariates) achieved the best accuracy with an RMSE value of 7.91 t/ha (about 20% of the mean). Results further showed that SOC stock correlated with general biome coverage, as the Grassland and Savanna biomes contributed the most (32.38% and 31.28%) to the overall SOC pool in South Africa. However, despite their smaller footprint, Forests (44.12 t/h) and the Indian Ocean Coastal Belt (43.05 t/h) biomes demonstrated the highest SOC sequestration capacity. The restoration of degraded biomes is advocated for, in order to boost SOC storage; but a balance between carbon sequestration capacity, biodiversity health, and the adequate provision of ecosystem services must be maintained. To this end, these findings provide a guideline to facilitate sustainable SOC stock management within South Africa's major biomes and indeed other regions of the world.

Keywords: Soil organic carbon; Biomes; Remote sensing; Climate; Topography; Deep learning

6.1. Introduction

A surge in extreme climatic events have forced global attention towards the need to fast-track climate change mitigation strategies (Roberts, 2021, Arletti *et al.,* 2021). Research has shown that the sequestration of carbon by biomass or soil could afford the international community time to address this problem (IPCC 2016, Baldock *et al.,* 2012). Soil organic carbon (SOC) offers the biggest terrestrial carbon pool and determines both the quantity and quality of soil ecosystem services at varying scales (Jiao *et al.,* 2020; Odebiri *et al.,* 2020b). Consequently, SOC stock management is of particular importance to global policymaking (IPCC, 2021). However, contextualizing the role of SOC amongst climate change mitigation frameworks requires a profound understanding of large-scale SOC pools and their interaction with both environmental and anthropogenic factors (Keskin *et al.,* 2019).

Biomes are large naturally occurring ecological zones made up of similar types of flora and fauna that can influence the global carbon cycle and whose distribution is dictated by distinct climatic envelopes (Schimel *et al.,* 2015). According to a 2009 assessment report by the United Nations Environment Programme (UNEP), over 2000 gigatonnes (Gt) of carbon are stored in the world's biomes, with the majority of this carbon accumulated in the soil (Trumper, 2009). Within South Africa’s unique biomes, SOC also constitutes the majority of the terrestrial carbon pool, with an estimated average sink of 6,396 gC/m^2 (Department of Environmental Affairs, 2017). However, anthropogenic activity has significantly affected SOC pools across the globe, with biomes losing approximately 116 Gt of SOC over the last 12,000 years (Amanuel *et al.,* 2018). Specifically, nearly 20% of South Africa's natural biomes have been transformed due to continuous land degradation, cultivation, and urbanization (Schulze and Scuttle, 2020; Venter *et al.,* 2017). Du Preez *et al.,* (2011) also noted that prolonged interference with South Africa’s biomes notably in the topsoil, reduces SOC stocks by around 45% on average; while Griscom *et al.,* (2017) suggests that about 2 Gt of South Africa’s SOC has been lost since 10,000 BCE.

Whereas, recent studies (e.g. Schulze and Scuttle, 2020; Venter *et al.,* 2021) have investigated the spatial distribution of SOC across South Africa; these studies were constrained to a comparison between two land-uses (natural vegetation and agricultural land use), and failed to explore the broader influences of biomes, their distinct carbon-feedbacks characteristics, and the possible impacts of land-use change upon South African SOC stock management. Importantly, the storage and persistence of SOC across different biomes are characterized by significant differences and are not only driven by intrinsic abiotic soil factors (such as topography, mineralogy and texture), but are also influenced by climate, plant density, edaphic conditions, and interference history (Wang *et al.*, 2021; Wieder *et al.*, 2018; Rutherford *et al.*, 2006). Yet, it is still uncertain which of these factors predominate in distinct biomes and at different geographical scales (Georgiou *et al.*, 2021). Therefore, continuous mapping and monitoring of the spatial variability of SOC across major biomes is vital to informing integrated policies and initiatives aimed at preserving existing stocks and reducing carbon emissions (Odebiri *et al.,* 2020b; Trumper 2009).

The adoption of geospatial techniques in digital soil mapping has attracted significant attention and has become a better alternative to the tedious traditional strategies of SOC determination (Odebiri *et al.,* 2021; Guo *et al.,* 2020; Odindi *et al*., 2016). This is facilitated by advancements in earth observation satellites (big data) in various volume, velocity, variety and veracity, which are capable of providing updated, consistent and spatially explicit assessment of SOC and its dynamics, particularly at a landscape scale (Xu *et al.,* 2019; Hamida *et al*., 2018; Kumar & Mutanga 2018). Moreover, the performance of these image datasets to effectively model SOC distribution is largely dependent upon the prepossessing procedures and the algorithm used (Padarian *et al.*, 2020; Wadoux *et al.*, 2019). Recent research comparing SOC accumulation from models to long-term field measurements have also indicated that both the models and observations still have considerable uncertainties, thus, requiring benchmarking analytical strategies to increase accuracy (Georgiou *et al.,* 2021; Wieder *et al.,* 2018; Sulman *et al.,* 2018). In addition, the ability of any analytical model to quantify the geographic variability of SOC levels across large swaths is contingent on its ability to contend with a high degree of dimensionality.

Use of deep learning (DL) algorithms have recently emerged as innovative analytical strategies to improve SOC mapping (Ma *et al.,* 2019). DL models are multi-layered representation-learning algorithms that uses nonlinear functions to extract information from lower to higher layers (Zhu *et al.,* 2019). In contrast to other geostatistical and conventional machine learning (ML) algorithms, DL frameworks can exploit feature representations exclusively learned from data which can significantly bolster mapping capabilities (Odebiri *et al.,* 2021). Furthermore, the intricate nonlinear nature of SOC and its relationship with the environment, can often present with a high degree of variability, which may become too complex for conventional ML algorithms, particularly at vast spatial extent (Ma *et al.,* 2019; Padarian *et al.*, 2019; Kumar *et al.,* 2016). Conversely, DL approaches can improve learning procedures by extracting relevant non-linear and complex attributes from data to improve results (Minh *et al.,* 2018). The intrinsic several hyper-parameters integrated within DL models also allows users to adjust training processes to produce more accurate results (Odebiri *et al.,* 2022). To this end, this study adopted the use of remotely-sensed-topo-climate datasets and a deep learning technique to estimate the spatial distribution of SOC across South Africa's major biomes.

6.2. Methodology

6.2.1. Brief description of South African Biomes

South Africa is home to a diverse array of biomes (Figure 6.1). According to Mucina and Rutherford (2006), South Africa is made up of nine distinct biomes including the Desert, Savanna, Forests, Fynbos, Succulent-Karoo, Nama-Karoo, Grassland, Albany Thicket and Indian Ocean Coastal Belt. A brief description of the biomes is given below. See Mucina and Rutherford (2006), for a detail description of South Africa's biomes.

6.2.1.1. Savanna

The Savanna biome covers about 399 600 km^2 of the country's surface and is categorized by an herbaceous layer dominated by grass species and a very open tree layer (Minasny *et al.,* 2017). Generally, savannas are found below altitudes of 1500 m, however, some parts have extended to 1800 m. Rainfall varies from 1350 mm at the higher altitudes to less than 200 mm. Savannas are mostly dominated by leptosols, with a high clay content and swelling properties (Venter 1990). Savannas can be considered species rich and are home to more than 34 large African herbivore species (Mucina and Rutherford 2006).

6.2.1.2. Grasslands

In South Africa, grasslands are limited to mainly flat to rolling topography between 300 - 400 m, but patches can also be found along the escarpment at greater altitudes. South African grasslands can be divided into either sourveld or sweetveld based upon moisture availability (Ellery *et al.,* 1955). Sourveld (moist grasslands) consist of leached and dystrophic soils and a high canopy coverage of sour grasses. Meanwhile, sweetveld (dry grasslands) are comprised of sweet grasses which grow upon eutrophic soils that are less leached (Ellery *et al.,* 1955). Sourveld are generally located at highly altitudes with a higher water supply, whilst sweetveld are found at lower altitudes. Grasslands are characterized by a high species richness and a wealth of endemic species peculiar to Southern Africa (Mucina and Rutherford 2006).

6.2.1.3. Nama-Karoo

Nama-Karoo covers almost 20 % of the country and extends from the western half of South Africa to it south-eastern border with Namibia. Nama-Karoo can be considered an arid-biome, with precipitation ranging from 70mm to 500mm. Rainfall generally occurs within the summer months, and temperatures range from -5° C in winter to 43° C in summer (Desmet & Cowling, 1999). Soils are considered base-rich and are weakly structured and skeletal. The biome comprises of plains dominated by low shrubs, C_3 and C_4 grasses mixed with geophytes, succulents and forbs (Mucina and Rutherford 2006). The region generally houses a variety of birdlife, ostrich (*Struthio camelus*) and springbok (*Antidorcas marsupialis*) (Rutherford *et al.,* 2006).

6.2.1.4. Fynbos

The Fynbos biome is characterized by evergreen small-leaved shrubs of the same name, which are dependent upon fire and are entirely endemic to South Africa (Mucina and Rutherford 2006). The biome, which forms part of the Cape Floristic Region and houses a large variety of endemic flora and fauna, occurs within relatively moist climates, across less than 7 percent of the country (Cox 2001). A large variation of soil types can be associated with the Fynbos region, such as heavy-textured soils, coastal plain soils, and silecrete and ferricrete rich soils amongst others (Mucina and Rutherford 2006).

6.2.1.5. Succulent-Karoo

The Succulent-Karoo Biome is the fourth largest biome in South Africa and covers 111 000 km. This Semi-desert region is characterized by a mild climate, which is prone to winter rainfall, with mean annual precipitation ranging from 100mm to 200mm (Mucina and Rutherford 2006). Succulent-Karoo soils are generally fine-grained, poorly leached with a high pH. These soils allow for an extremely species rich habitat and forms a home to a high diversity of distinct dwarf leaf-succulent shrubs (Mucina and Rutherford 2006).

6.2.1.6. Albany Thicket

The Albany Thicket biome consists of various vegetation types, including: Spekboomveld, Noorsveld, Valley bushveld, and False Karroid Broken veld located along the semi-arid regions of the Eastern and Western Cape. The biome provides a high diversity of plant species, such as stem and leaf succulents, geophytes, grasses, and semi-deciduous woody shrubs (Mucina and Rutherford 2006). Annual precipitation ranges from 200 to 950mm. Although thicket soils are not limited to any particular soil type, soils have been found to be high in carbon, calcium and potassium due to unique below-ground animal activity (Mucina and Rutherford 2006).

6.2.1.7. Indian Ocean Coastal Belt

The Indian Ocean Coastal Belt Biome occurs along an 800 km strip of South Africa’s Eastern coastline along altitudes of 0 to 600 m. An annual rainfall of between 820mm and 1270 mm mostly occurs during the summer months, with a mean annual temperature of 22° C. The biome is open to dense savanna vegetation, intermixed with vast areas of forest and sourveld grassland (Mucina and Rutherford 2006).

6.2.1.8. Desert

South African Deserts are hyper-arid regions which are home to a high diversity of organisms which are adapted to arid environments. The eastern parts are dominated by a variety of drought tolerant grasses and wood-shrubs, while the western parts contain leaf-succulent chamaephytes (Mucina and Rutherford 2006). Temperatures may reach 47.8° C, while mean annual rainfall remains below 70mm. The alluvial soils are slow forming and may be subject to erosion processes (Mucina and Rutherford 2006).

6.2.1.9. Forests

The Forest biome is predominately scattered along the Eastern and Southern regions and covers a mere 7% of the country’s landmass (Mucina and Rutherford 2006). Indigenous forests are defined as multi-layer vegetative stands which are comprised of evergreen trees with a crown cover larger than 75%, and a stand height ranging from 3 to 30m that are endemic to South Africa. They persist in regions with a mean annual rainfall greater than 725 mm. Soils derived from the sandstone, shale and dolomite geology, vary in depth, and nutrient status (Mucina and Rutherford 2006).

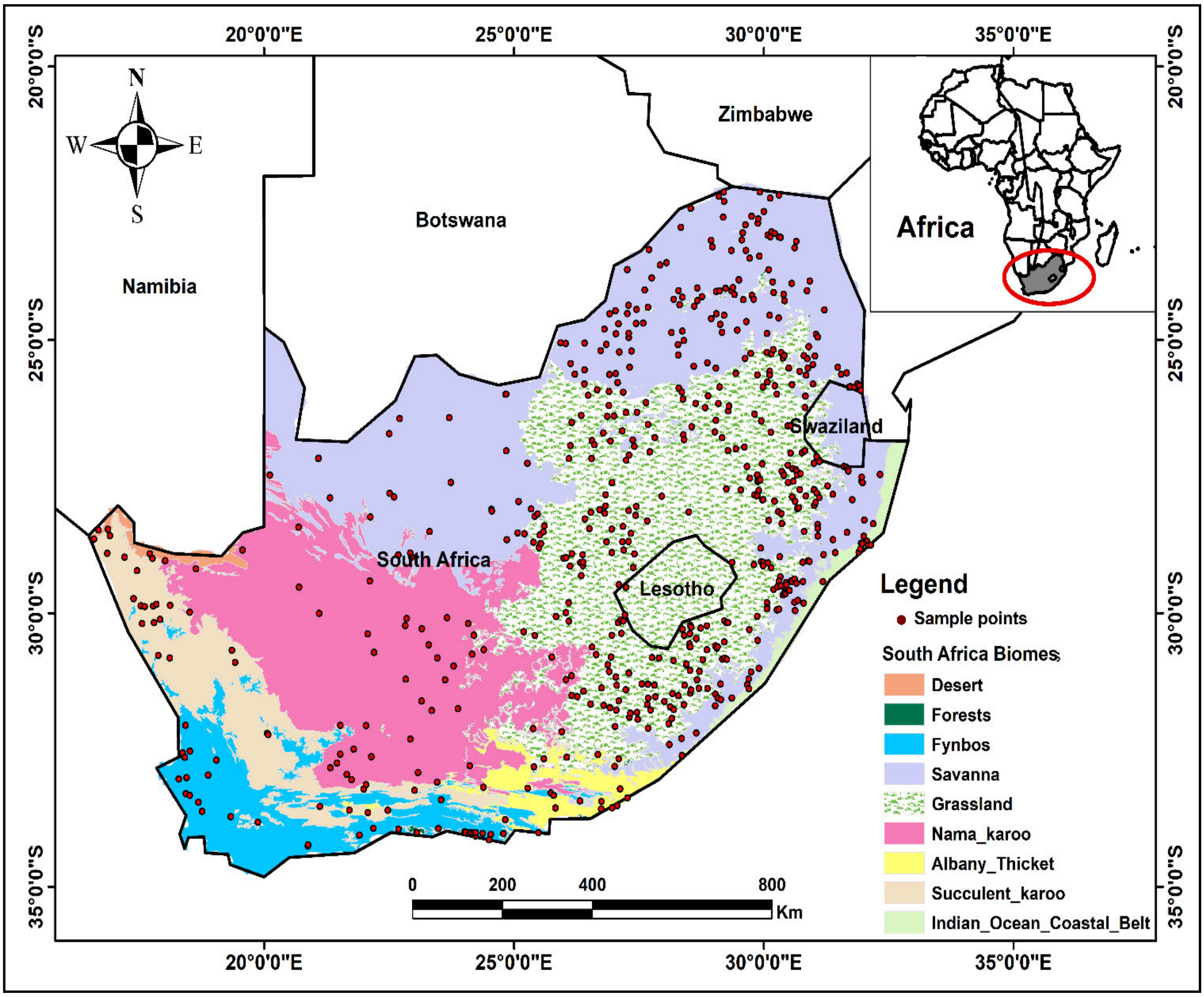


Figure 6.1. South Africa's location and the spatial distribution of soil samples spread across major biomes

6.2.2. Soil profile data

Majority of the soil profile data used for this research was acquired from the International Soil Reference and Information Centre (ISRIC), while the rest of the soil profiles were acquired from previous soil research projects conducted at the University of KwaZulu-Natal's School of Agricultural, Earth and Environmental Sciences (SAEES). ISRIC is a non-profit scientific organization that collaborates with various countries, including South Africa, to produce high-quality data on various soil physical and chemical properties on a global scale. The latest release of ISRIC soil records (https://www.isric.org/) was last updated in 2020 and comprises over 150,000 observations from 173 nations (Batjes *et al.*, 2020). We acknowledge that the bulk of these sample points differ in terms of collection location and time, and the methodologies for determining SOC carbon concentration for the samples vary per nation

(Hengl *et al.*, 2017). However, ISRIC (Batjes *et al.*, 2020; Batjes *et al.*, 2017) has developed a comprehensive approach to ensure uniformity and accessibility of input soil profile data (https://www.isric.org/explore/wosis/accessing-wosis-derived-datasets). Consequently, point observations encompassing South Africa, with their corresponding SOC content, as well as the bulk density, was retrieved from ISRIC records together with the accessible SAEES data. Pearson *et al.,* (2007) formula was used to determine SOC stocks at each point as indicated below;

$$SOC\ accumulation\ (t/h) = SOC\ concemtration\ x\ volume\ density\ x\ soil\ depth \qquad (1)$$

6.2.3. Acquisition of image data

6.2.3.1. Sentinel 3

The European Space Agency's (ESA) recently launched Sentinel-3 Ocean and Land Colour Instrument (OLCI) image data was utilized in this study. Sentinel-3 consists of 21 spectral bands between 400 and 1020 nanometres (Vaudour *et al.*, 2019). Its 300-meter spatial resolution, 1270-kilometer geographical coverage, and less than 2-day return interval enables reliable, rapid, and continuous mapping at landscape scales (Li and Roy, 2017). Although Sentinel-3 possesses spectral bands critical for soil and vegetation monitoring, its capabilities are still relatively underexplored within the field of digital soil mapping (Li *et al.,* 2021; Kokhanovsky *et al.*, 2019). Four image tiles with less than 10% cloud coverage were downloaded between March and April 2021 from ESA (https://scihub.copernicus.eu/). Utilizing the Sentinel Application Platform (SNAP v. 6.0), the downloaded images were geometrically, radiometrically and atmospherically corrected. Subsequently, the pre-processed tiles were mosaicked into a single tile, and clipped to the South African border. Vegetation indices have been shown to be beneficial in explaining SOC fluctuation and distribution in several studies (Wang *et al.*, 2021; Odebiri *et al.,* 2020a; Guo *et al.*, 2020). As a result, eleven relevant vegetation indices were created using various Sentinel-3 spectral band combinations (Table 6.1).

6.2.3.2. Topo-climate metrics

Previous research has highlighted topographic and climatic factors as important determinants of SOC dispersion (Li *et al.*, 2018; Fissore *et al.*, 2017; Wang *et al.*, 2012). Spatial topography metrics are divided into three classes, according to Li *et al.*, (2018); these include local, non-local, and mixed topographical metrics. Local metrics including slope, elevation, and curvatures evaluates surface geometry at a specific location, whereas non-local metrics including relief, catchment area, openness, and flow build-up depict the relative locations of selected points (Li *et al.,* 2018). The topographic wetness index (TWI), slope length factor and stream power index, are examples of mixed factors which integrate both local and non-local topographic elements (Li *et al.,* 2018; Florinsky 2016; Lang *et al.,* 2013). Nineteen different terrain variables (Table 6.1) were chosen for this study that span across the three classifications (local, non-local, and mixed) which were generated from a Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) using the SAGA GIS (2.3.2) and ArcGIS Pro 2.8 software. In addition, the mean temperature and rainfall for South Africa were computed using the WorldClim datasets with one square kilometre (1km^2) 30 arc-seconds spatial resolution (http://www.worlclim.org/). Grids of diverse climatic conditions spanning over 30 years such as the average yearly temperature and rainfall together with the wettest, driest, coldest, and hottest quarters as well as months of the year, can be obtained from the WorldClim database. Both the DEM and WorldClim datasets were resampled to correspond to the spatial resolution of the Sentinel 3 data (300m). This was done using the raster resample function within the ArcGIS Pro 2.8 environment. The justification for use of the vegetation indices and the topo-climate variables in this study are detailed in Odebiri *et al.,* (2020a) and Odebiri *et al.,* (2020b).

Table 6.1. Remotely sensed-topo-climate variables

Metrics	**Description/formula**	**Reference**
Remotely sensed data		
All 21 bands of Sentinel-3	—	ESA
Normalized Difference Vegetation Index(NDVI)	$\frac{NIR - RED}{NIR + RED}$	Rouse (1974)
Soil Adjusted Vegetation Index (SAVI)	$\frac{NIR - RED}{NIR + RED + 0.5}(1 + 0.5)$	Huete (1988)
Optimized Soil Adjusted Vegetation Index (OSAVI)	$(1 + 0.16)\ (NIR - RED)/(NIR + RED + 0.16)$	Jamalabad and Abkar , (2004)
Modified Soil Adjusted Vegetation Index (MSAVI)	$\frac{2NIR + 1 - \sqrt{(2NIR + 1)^2 - 8(NIR - RED)}}{2}$	Qi (1994)
Enhanced Vegetation Index (EVI)	$2.5 \times \frac{NIR - RED}{(\mathrm{NIR} + 6 \times \mathrm{RED} - 7.5\ \times \mathrm{BLUE} + 1)}$	Huete (1999)
Ratio Vegetation Index (RVI)	$\frac{NIR}{RED}$	Baret (1991)

Renormalized Difference Vegetation Index (RDVI)	$\frac{(NIR - RED)}{(NIR + RED)^{1}/2}$	Roujean and Breon (1995)
Transformed Vegetation Index (TVI)	$\sqrt{(NDVI)} + 0.5$	Deering (1975)
Difference Vegetation Index (DVI)	$NIR - RED$	Richardson (1977)
Green Normalized Difference Vegetation Index (GNDVI)	$\frac{NIR - GREEN}{NIR + GREEN}$	Gitelson (1998)
Topography		
Elevation (DEM)	Ground height	Davy and Koen (2014)
Slope	The steepness of the ground	Li *et al.,* (2014)
Aspect	Slope direction	Rezaei and Gilkes, (2005)
Topographic wetness index (TWI)	Steady state wetness index	Lang *et al.,* (2013)
Topographic position index (TPI)	Areas that are greater or lesser than the mean of their scenery	Wiesmeier *et al.,* (2011)
General curvature (Gen Curv)	Curvature both horizontally and vertically	Li *et al.,* (2014)
Valley Depth	Relative Heights	Böhner *et al.,* (2001)
Standardized Height (SH)	Position of relative height and slope	Böhner *et al.,* (2001)
Normalized Height (NH)	Position of relative height and slope	Böhner *et al.,* (2001)
Positive Openness (PO)	Drainage features, soil water content	Seijmonsbergen *et al.,* (2011)
Negative Openness (NO)	Drainage features, soil water content	Seijmonsbergen *et al.,* (2011)
Direct Insolation (Dir Ins)	Potential Incoming insolation	Rodriguez *et al.,* (2002)
Diffuse Insolation (Dif Ins)	Solar radiation	—
Catchment Area (Catch A)	Runoff velocity and volume	Kasai *et al.,* (2001)
Convergence index (Con Idx)	Shows relief structure as a set of convergent (channels) and divergent (ridges) areas.	Ließ *et al.,* (2016)
Sky view factor (SVF)	Visibility	Ließ *et al.,* (2016)
LS factor	The influence of slope length on erosion	Böhner &Selige (2006)
Plan curvature (Plan Curv)	Horizontal (contour) curvature	Troch *et al.,* (2002)
Profile curvature (Pro Curv)	The pace at which the slope changes vertically	Ritchie *et al.,* (2007)
Climate data		
Rainfall	Mean annual precipitation	Odebiri *et al.,* (2020b)
Temperature	Mean annual temperature	Odebiri *et al.,* (2020b)

6.2.4. Concrete Autoencoder-Deep neural network (CAE-DNN)

A model performance may be significantly influenced by the heterogeneous nature of remotely sensed data (Masemola and Cho 2019). As such, to facilitate dimensionality reduction and improve model accuracy, this study adopted a hybrid deep learning (DL) framework based on

Concrete Autoencoder-Deep neural network (CAE-DNN) for SOC retrieval. Firstly, the framework utilized Concrete Autoencoders (CAE) for variable selection and the removal of redundant variables. Thereafter, a Deep neural network model was applied for the SOC prediction process.

The concept of CAE (Figure 6.2) was initially proposed by Abid *et al.,* (2019). CAE is an end-to-end distinguishable unsupervised approach for global variable selection that efficiently finds a subset of the most relevant features while simultaneously learning a neural network. CAE is a dimensionality reduction adaptation of the standard autoencoder that consists of two basic components: the encoder and the decoder. However, instead of employing fully connected layers for the encoder, a concrete selector layer with a user-defined number of nodes is introduced which chooses stochastic linear combinations of input features during training and converges to a discrete set of features by the conclusion of training (Abid *et al.,* 2019). The concrete selector layer is made up of concrete arbitrary variables and is controlled by a temperature parameter (Jang *et al.,* 2016). The temperature of the concrete selector layer is gradually reduced during the training phase, encouraging the learning of a user-specified number of discrete features (Maddison *et al.,* 2016). Its architecture consists of a single encoding layer and random decoding layers. The CAE has proven to be a powerful dimensionality reduction framework and generally performs better in optimal variable selection when compared to other conventional variable selection strategies (a detailed description of CAE can be found in Abid *et al.,* 2019).

To analyse the efficacy of the CAE variable selection strategy, we introduced the Boruta feature selection algorithm as a comparison. Boruta, which is based on the random forest (RF) concept, creates shadow features for given datasets, trains an RF classifier on the datasets, and assigns an essential measure score to each variable, indicating the ones to be selected and the ones to be discarded (See Kursa and Rudnicki, 2010 for a detailed Boruta description).

In recent years, deep neural networks (DNN) have piqued the interest of the remote sensing community (Odebiri *et al.,* 2021). The DNN is a useful and dependable approximate model for determining linear and complex interrelations involving the target and other covariates (Odebiri *et al.,* 2022; Wang *et al.,* 2020). It comprises of three major layers including the input, hidden, and output, with many neurons within them. Until the final predicted (output) neuron, neurons in one layer are connected to neurons in the following layer. The DNN is generally based on the multilayer perceptron (MLP) architecture, but differs from the standard MLP models in that it incorporates several hidden layers and hyper-parameters (Taghizadeh-Mehrjardi *et al.,* 2020). Nevertheless, extra caution should be exercised when calibrating the DNN architectures to elude overfitting (Liu *et al.*, 2018). In such scenarios, a dropout regularization technique can be performed on each of the hidden layers' nodes, which are then aggregated for estimation. For a given input layer (vector ***X***) with a ***L*** hidden layer and an output layer (vector ***Y***), the following is a basic mathematical representation of DNN framework: (Wang *et al.,* 2020);

$$\mathrm{Z}^1 = \sigma_1(\mathrm{W}^1\mathrm{X} + \mathrm{b}^1), \mathrm{Z}^2 = \sigma_2(\mathrm{W}^2 z^1 + \mathrm{b}^2)$$
$$\mathrm{Z}^\mathrm{L} = \sigma_\mathrm{L}(\mathrm{w}^\mathrm{L} z^{L-1} + \mathrm{b}^\mathrm{L}),\ \mathrm{Y} = \ \mathrm{W}^{\mathrm{l}+1}\mathrm{Z}^\mathrm{L} + \mathrm{b}^\mathrm{L}, \theta = [\mathrm{W}^\mathrm{i}, \mathrm{b}^\mathrm{i}]_{\mathrm{i}=1}^{\mathrm{L}+1} \tag{2}$$

Where *Wi* and bi are the ith layer's weights and biases, accordingly. *L* + 1 denotes the output layer (i.e. *Y* = *N* (*X*; *Θ*)) and σi denotes the ith layer's activation function.

The mean squared error (MSE) can be used to examine the loss function *L* of the output and input variables, which is represented as follows;

$$\mathrm{MSE_{DATA}} = \mathrm{L}(\theta) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} |\mathrm{NN}(\mathrm{X_i};\ \theta) - \mathrm{Y_i}|^2 \tag{3}$$

Where *N* represents the total number of labelled data. An optimization algorithm can be used to minimize the loss function *L*.

Figure 6.2 and Table 6.2 depict the CAE-DNN architecture schematic, and the defined range and selected sets of hyper-parameters tested for best output, respectively. Following the CAE feature selection procedure, the SOC model was built using a total of 26 selected input variables. For training and testing, the full data was divided into ten equal sections and used in that sequence. The DNN was calibrated ten times, ensuring that every data point was used as validation. We employed the Bayesian Optimization approach with 10-fold cross-validation for hyper-parameter adjustment. To get the best results, 500 epochs, 6 hidden layers, the "adam" optimizer, and the LeakyReLU activation function were utilized. The CAE-DNN results were compared to the Boruta-DNN and a DNN model without variable selection. All the analysis was conducted within the python 3.8 API, except for the Boruta variable selection method that was conducted in R studio (Version 4.0.2)

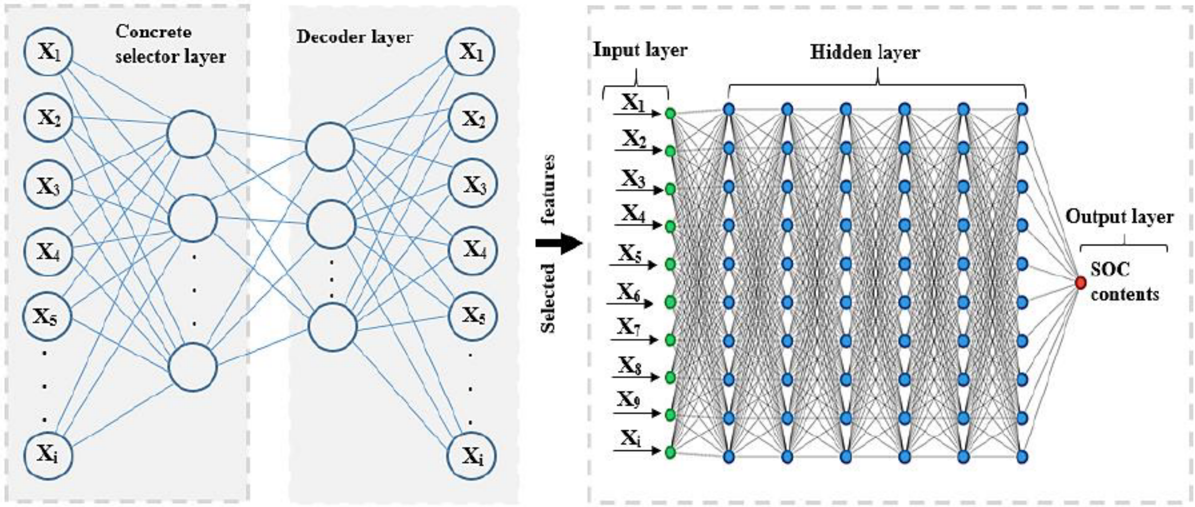


Figure 6.2. A schematic representation of the Concrete Autoencoder-Deep neural network (CA-DNN) architecture.

Table 6.2. Hyper-parameter tuning via Bayesian optimization

Hyper-parameters	Description	Range	Selected
Hidden layer	Hidden layers count	3–20	6
Neuron size	Count of nodes in the hidden layers	20–400	256
Network weight initialization	Initialized layers weights from input to output	uniform/he-normal	He-normal
Learning rate	Readjusts the weights of the network	0.001–0.05	0.01
Dropout regularization	Discarded neurons or nodes to reduce overfitting	0.2–0.5	0.3
Optimizer	Changes the attributes of a network in order to reduce the losses	Adam, RMSprop, SGD, Adamax, and Adadelta	Adam
Activation function	Specifies how a network node or network nodes in a layer converts the weighted sum of the input into an output.	Relu, LeakyReLU, Sigmoid, Tanh, and Softplus	LeakyReLU
Batch size	Number of training samples processed in one iteration	4 –100	32
Epocs	Number of training iteration	100-1000	500

6.2.5. Model evaluation metrics

Root mean squared error (RMSE), coefficient of determination (R^2), and Lin's concordance correlation coefficient (LCCC) is used to examine the fitting and generalization of the models created in this study, which are represented as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_{O,i} - X_{P,i}\right)^2}{n}} \tag{4}$$

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} (X_O - p)^2}{\sum_{i=1}^{n} (X_O - O')^2}\right] \tag{5}$$

$$\text{LCCC} = \frac{2\, r\, \sigma_O \sigma_P}{\sigma_P^2 + \sigma_P^2 + [O' - P']^2} \tag{6}$$

In this case *n* denotes the count of samples and X_O and X_P denote the actual and estimated SOC values, respectively. The means of the actual and estimated SOC are symbolized by *O'* and *P'*, respectively, while the variation of the actual and estimated value is expressed by $\boldsymbol{\sigma}_o$ and $\boldsymbol{\sigma}_p$. To eliminate sampling bias, the data is divided into ten uniform sets and successively transmitted into both the train and test datasets, ensuring that every set was equally utilized. In general, the most fitting model has a greater R^2 and LCCC as well as a smaller RMSE. In addition, the relevance of covariates was assessed in order to establish their importance to the model's accuracy. Generally, DL approaches frequently struggle to attain interpretability because they are incapable of automatically quantifying the impact of covariates during or after every task, which is why they are referred to as "black boxes" (Padarian *et al.*, 2019). Various ways have recently been developed to aid users in interpreting DL simulations (Pentos, 2016). One of these solutions is the SHApely Additive exPlanations (SHAP) strategy, which was implemented in this study. The primary idea behind SHAP is to allocate a unique mean value to variables in order to show the degree to which each variable contributed to the model's performance. The higher the average SHAP value for a particular variable the higher its influence on the models outcome. SHAP has a package for any type of ML algorithm, including a "DeepExplainer" function for DL models, and has several advantages over other techniques, including global and local interpretability.

6.2.6. Uncertainty quantification

Describing uncertainty and measuring the strength of research findings are critical in decision-oriented studies (Hamel and Bryant, 2017). Consequently, utilizing the typical ±1.64 standard deviation (SD) with a 90% confidence interval (C1) significance level, the upper and lower limits for the SOC maps produced by the CAE-DNN framework were developed in this work (Odebiri *et al.*, 2022). We used a 10-fold cross-validation under the premise that the model follows a normal distribution for each raster pixel (Emadi *et al.,* 2020). Thereafter, the estimated mean value of each pixel, as well as the 5th and 95th percentiles, were then retrieved. Finally, a distribution map for the computed mean, lower (5%) and higher (95%) confidence levels were created.

6.3. Results

6.3.1. Synopsis statistics

From the 1936 total soil samples used, SOC values ranged from 5.3 t/ha to 149 t/ha, with a mean and standard deviation of 39.8 t/ha and 17.3 t/ha, accordingly. A relatively high SOC variance of 43 percent was noted. The data diverged from the normal distribution curve, resulting in high skewness (1.9) and kurtosis (5.2) (Hair *et al.,* 2016). To address this, a natural logarithm transformation technique was used to enhance the data distribution, hence new values for skewness (0.41) and kurtosis (0.68) were obtained. After predictive analysis, the rescaled data was retransformed back to its initial scale.

6.3.2. Feature selection

The development of predictive models often requires the computation of the smallest possible number of the most influential variables (Odebiri *et al.,* 2020a). Consequently, a variable selection technique is often necessary to improve accuracy. Through an adjustment of the concrete selector layer of the CAE model, a significant reduction in dimensionality of the covariates was achieved (Figure 6.3), with the system identifying 26 out of 52 features for prediction. The selected features were Rainfall, NDVI, Band 8, EVI, Band 11, Elevation, Slope, Temperature, Band 7, TWI, SAVI, Band 12, Band 13, Band 9, DVI, Band17, General curvature, GNDVI, RVI, Band 10, Band 19, Band 18, Profile curvature, Band 6, Catchment area and Band 2.

Meanwhile, the Boruta selection approach (Figure 6.4) identified 28 important features, rejected 21 features, and flagged 3 features as tentative (where the algorithm is uncertain on the importance of the variable). The 28 selected variables by the Boruta also included all the variables selected by the CAE model together with Band 16 and Valley depth. The tentative features (n = 3) including Band 4, Aspect and Direct Insolation, were added to the important features after applying the “TentativeRoughFix” function within the R programming API, resulting in total of 31 predictive features used for the development of the Boruta-DNN model. Thereafter, the selected variables identified by CAE and Boruta were then used to build separate models and their results were compared to a DNN model that included all 52 variables.

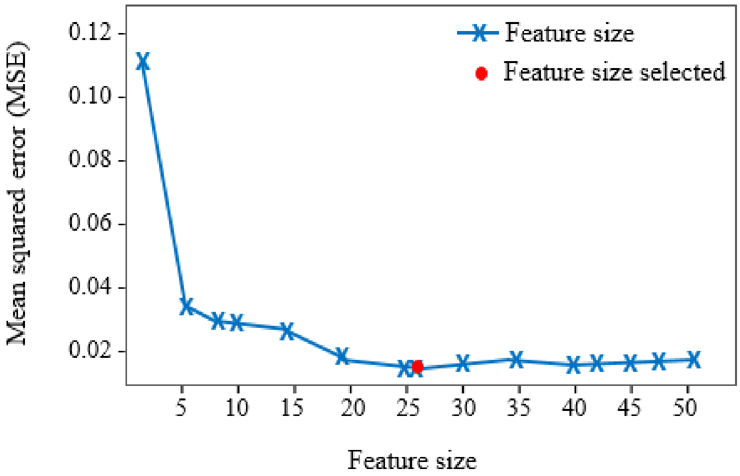


Figure 6.3. Scree plot of Concrete Autoencoder (CAE) showing mean squared error for different feature sizes. The red dot signifies the feature selected (*n* = 26)

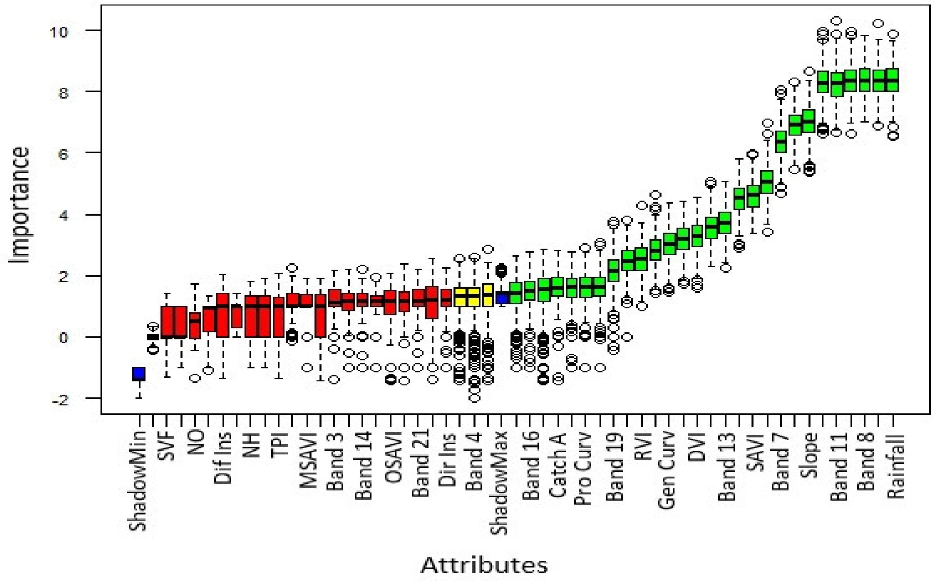


Figure 6.4. Variable selection (*n* = 31) by the Boruta algorithm; green bars are the selected; red bars are the rejected and yellow bars are the tentative features. Note that half of the selected and rejected features are not listed on the figure's x-axis due to space

6.3.3. Evaluation and performances of models

Table 6.3 shows the means and standard deviations estimates of the three models: CAE-DNN, Boruta-DNN, and a DNN model without a variable selection strategy. Figure 6.5 (ABC) also illustrates the relationship involving the actual and predicted SOC for each of the developed models (CAE-DNN, Boruta-DNN, and a DNN model without a variable selection strategy). The CAE-DNN model with the smallest number of features (*n* = 26) generated the best results, with RMSE, R^2, and LCCC scores of 7.9 t/h (19.88% of the mean), 69.94, and 85.35, respectively. Next is the Boruta-DNN model (*n* = 31), with an RMSE of 9.48 (23.83 % of the mean), R^2 of 68.14, and LCCC of 81.17. The DNN model with no variable selection (*n* = 52) had the poorest results (RMSE=10.71—26.92 % of the mean, R^2 = 66.7, and LCCC = 79.66). This suggests that having too many redundant variables in a model can considerably impact accuracy. The variable importance (Figure 6.5D) of the best performing model (CAE-DNN, RMSE = 7.9 t/h) was then used to examine the variables that best explain the distribution of SOC stocks across South African biomes.

Interestingly, all of the data categories used, including remotely sensed data derivatives, topography, and climate, had a considerable impact on the model output. The SHAP graph revealed that Rainfall, NDVI, Band 8, Elevation, Temperature, EVI, Band 11, Slope, RVI, and TWI were the top ten explanatory variables for SOC variability in the area.

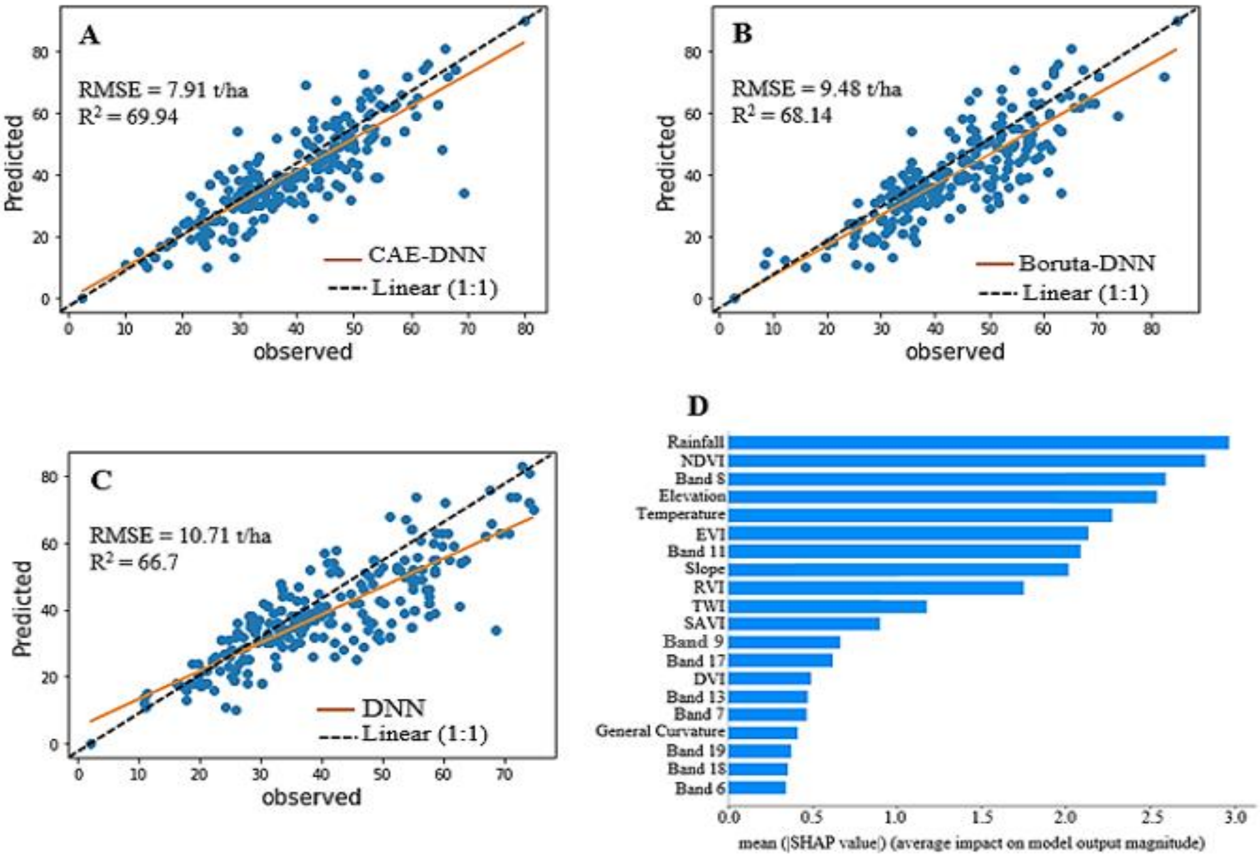


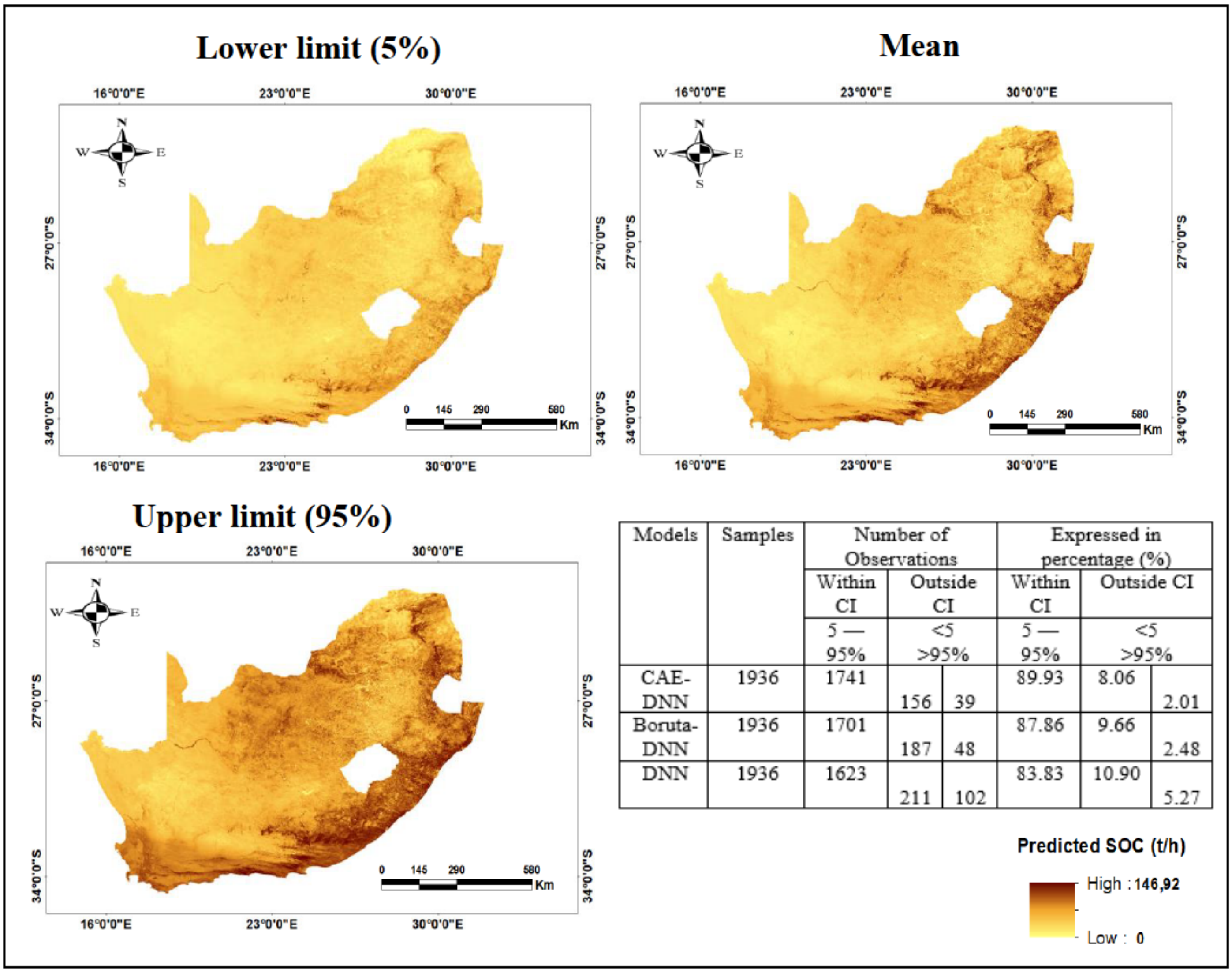
Figure 6.5. Predicted versus observed soil organic carbon (SOC) based on three models and variable importance measure: (A) Concrete Autoencoder-Deep neural network (CAE-DNN); (B) Boruta-Deep neural network (Boruta-DNN); (C) Deep neural network (DNN) without feature selection; (D) rank of covariates using SHAP method

Table 6.3. Output of Concrete Autoencoder-Deep neural network (CAE-DNN), Boruta-Deep neural network (Boruta-DNN), and Deep neural network (DNN) without feature selection. R^2: coefficient of determination; LCCC: lin's concordance correlation coefficient; RMSE: root mean square error (mean ±standard deviation).

Model	RMSE (t/ha)	R^2	LCCC
CAE-DNN	7.91±4.30	69.94±4.46	85.35±5.02
Boruta-DNN	9.48±5.06	68.14±4.92	81.17±6.07
DNN	10.71±5.21	66.7±6.02	79.66±5.03

6.3.4. SOC spatial estimation and uncertainty quantification

Figure 6.6 shows the variability and uncertainty of the estimated SOC for South Africa at the higher (95%), mean, and lower (5%) confidence intervals for the best model (CAE-DNN). The figures demonstrate a strong consistency from the top to lower bounds, with higher SOC stocks in highly vegetated regions compared to semi-vegetated or non-vegetated areas. For instance, the northern, eastern and southern parts of the mean map with darker colours indicating high SOC stocks are dominated by the savanna biome, which consist of dense natural forests and plantations as well as the grassland biome with rich species capable of sequestering enormous amounts of carbon. The table within Figure 6.6 shows the count and percentage (%) of SOC that fall between the specified 90% prediction intervals using the various models with 10-fold cross-validation. According to Minasny *et al.,* (2016), 90 percent of the samples are expected, in theory, to fall between the given interval. The CAE-DNN model yielded the best consistent uncertainty, with roughly 90% and 10% of the observations falling inside and outside the prescribed CI, respectively. The Boruta-DNN model closely followed, with over 87 and less than 15 percent of samples falling inside and outside the stated CI, respectively. The DNN model without variable selection technique also produced a reliable uncertainty estimates indicating that about 84% of the samples fall within the CI. It is however pertinent to emphasize that the uncertainty quantification conducted in this study was exclusively based on the models and their parameters, and not the spatial uncertainty of the raw data used. As such, our uncertainty quantification may not be entirely representative of the actual situation on ground.



Models	Samples	Number of Observations			Expressed in percentage (%)		
		Within CI	Outside CI		Within CI	Outside CI	
		5 — 95%	<5	>95%	5 — 95%	<5	>95%
CAE-DNN	1936	1741	156	39	89.93	8.06	2.01
Boruta-DNN	1936	1701	187	48	87.86	9.66	2.48
DNN	1936	1623	211	102	83.83	10.90	5.27

Figure 6.6. Spatial estimation of soil organic carbon (SOC) at lower limit (5%), mean, and upper bound (95%), derived utilizing remotely sensed-topo-climate data and Concrete Autoencoder-Deep Neural Networks (CAE-DNN). The table within provides the proportion of samples falling inside the prescribed CI (i.e. 5 and 95 percent) for the models

6.3.5. SOC Spatial distribution across various South Africa biomes

The statistical overview as well as the spatial distribution of SOC throughout South Africa's distinct biomes are highlighted in Table 6.4 and Figure 6.7. The Savanna biome is the largest, covering about 33% of the entire land area. Grassland (28.41 %), Nama Karoo (20.53 %), Succulent Karoo (6.86 %), Fynbos (6.73 %), Albany Thicket (2.48 %), Indian Ocean Coastal Belt (1.31 %), Desert (0.58 %), and Forests (0.08 %) are the other biomes in order of size. Despite being the second largest in area, the Grassland biome accumulates the most SOC stocks (32.38%), while the Forests biome accumulates the least (0.14 % stocks). The Grassland biome also has the maximum concentration per single soil profile (146.92 t/h), while the Savanna biome has the minimum with 6.48 t/h. In addition, the mean values of the SOC stocks for each biome type provides an insight of the carbon sequestration capacity of each biome in respect to their total land area. To this end, the Forests biome, though the smallest in size, has the highest mean value of 44.12 t/h, indicating a higher sequestration potential compared to other biomes. The Indian Ocean Coastal Belt biome also has a high mean value (43.05 t/h) close to that of the Forests, while the Albany Thicket, the Fynbos and the Grassland biomes all possess a relatively high mean values of 38.76 t/h, 36.20 t/h and 35.77 t/h, respectively. Unsurprisingly, the desert biome has the smallest mean value of 14.02 t/h; indicating a low SOC retention capability. It is spatially evident that the majority of South Africa's SOC stocks are mostly distributed along the east coast of the country (Figure 6.6). Moreover, biomes with higher average SOC concentrations such as Grassland, Savanna, Forests, Indian Ocean Coastal Belt, Albany Thicket and parts of the Fynbos are prevalent within KwaZulu-Natal, Eastern Cape, and Mpumalanga provinces (Figure 6.7). Conversely, lower SOC stocks for other biomes are mostly found within the desert, arid and semi-arid Western and Southern parts of the country.

Table 6.4. A summary statistics of the spatial distribution of SOC stocks across South Africa biomes. The 5th and 95th percentile values are in square parentheses; lowest and highest values for each column is emboldened.

Biome	Area (%)	Total SOC (%)	Min SOC (t/h)	Mean SOC (t/h)	Max SOC (t/h)
Savanna	**33.02**	31.28	**6.48**[2.7, 10.68]	28.28[23.9, 33.4]	86.88[81.4, 91.2]
Grassland	28.41	**32.38**	10.46[5.81, 14.9]	35.77[30, 39.91]	**146.92**[141, 151.]
Nama Karoo	20.53	15.81	10.02[6.1, 15.21]	21.51[17.8, 26.5]	63.18[58.9, 68.7]
Fynbos	6.73	8.97	10.69[5.84, 15.7]	36.20[31.1, 40.5]	59.78[54.3, 64.2]
Succulent Karoo	6.86	5.61	6.87[2.9, 11.03]	22.28[17.9, 27.2]	45.42 [40.8, 50.3]
Albany Thicket	2.48	3.58	20.02[14.95, 25]	38.76[32.7, 42.8]	64.83[59.1, 69.8]
Indian Ocean Coastal Belt	1.31	1.95	7.28[3.12, 12.1]	43.05[39.8, 48.7]	68.94[63.6, 73.1]
Desert	0.58	0.29	6.56[2.5, 11.34]	**14.02**[9.91, 19]	**18.62**[13.6, 23.1]
Forests	**0.08**	**0.14**	**36.82**[31.2, 40.6]	**44.12**[39.9, 49.1]	66.71[61.7, 70.4]

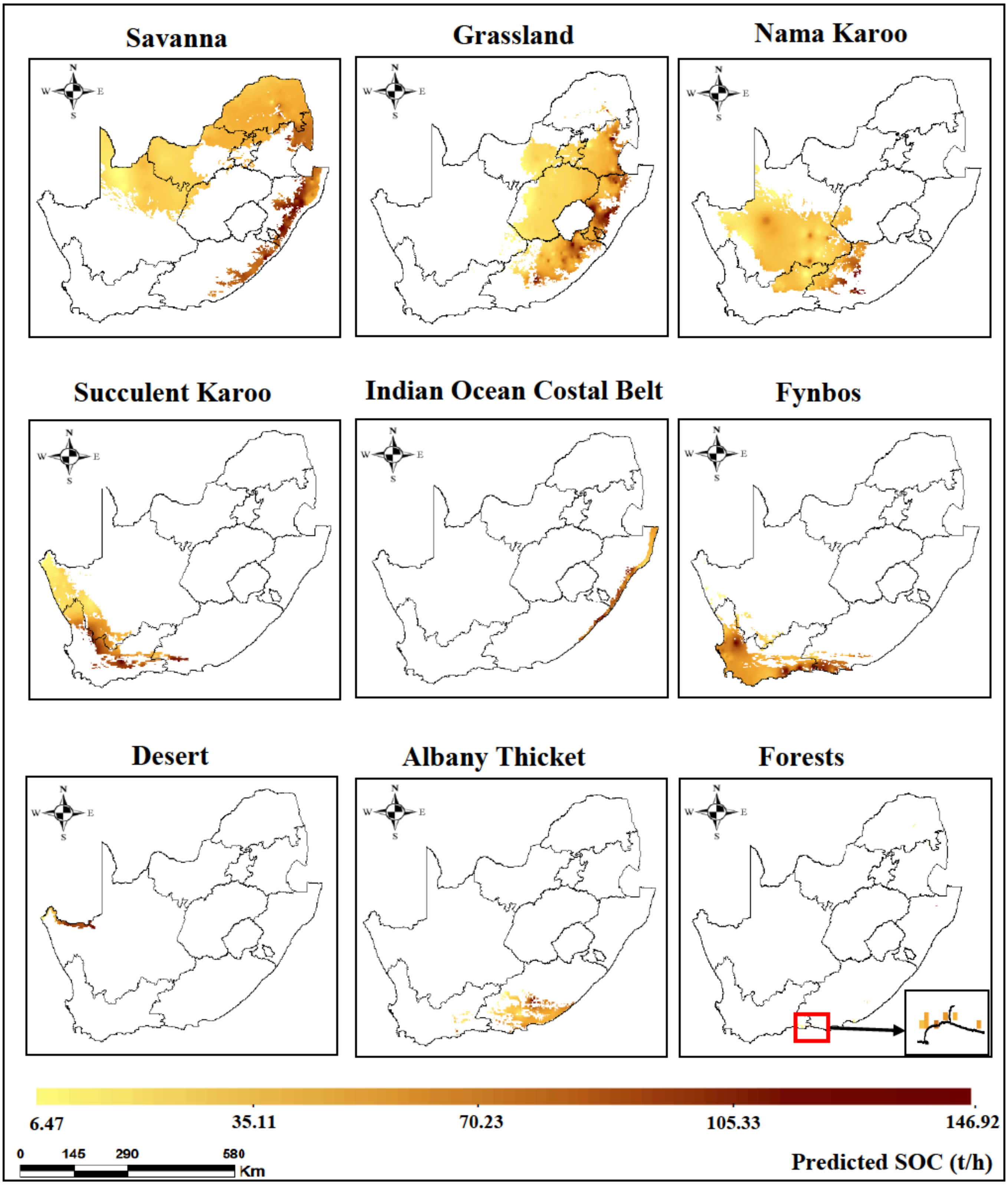


Figure 6.7. Soil organic carbon (SOC) stocks distribution across South Africa's biomes; The Forests biome is barely visible at full extent; hence, a zoomed-in section in the red box is depicted within the adjacent box pointed by the black arrow

6.4. Discussion

6.4.1. Assessing the distribution of soil organic carbon stocks within South Africa biomes

Understanding the spatial distribution of SOC across different biomes is vital for large-scale carbon management and planning (De Deyn *et al.,* 2008). The distribution of SOC stocks, however, are often dependent on an ecosystem type and its exposure to specific topographic and climatic variables (Lal, 2009). In light of this, we mapped the spatial distribution of SOC across South Africa's major biomes using a combination of remotely-sensed-topo-climatic datasets and Concrete Autoencoder-Deep Neural Networks (RMSE = 7.91 t/h). The resultant model showed that almost two thirds of South Africa's SOC stock is currently stored beneath the Grassland (32.38 %) and Savanna (31.28 %,) biomes (Figure 6.7). This corresponds with the findings of Venter *et al.,* (2021), who found that the vast expanses of the Savanna (404, 757 km^2) and Grassland (269, 920 km^2) biomes contributed significantly towards their overall SOC accumulation.

South African grasslands are home to a variety of diverse grasses, herbaceous graminoids and forbs (Egoh *et al.,* 2011). This enables a high basal cover that facilitates the decomposition of organic litter, which in turn leads to increased soil organic matter and greater SOC production (Mills and Fey, 2003). The highest concentrations of SOC were predominant in the sub-escarpment grasslands along the eastern seaboard (Figure 6.7). Here, long-lived sourveld grasses with deep underground storage organs aide in the accumulation of SOC. Sourveld grasses, such as the KwaZulu-Natal Sandstone Sourveld, may directly contribute towards carbon stocks through elevated levels of woody biomass. Although these grasslands are prone to fire, the majority of its carbon is stored underground (i.e., roots and soil), making them more resilient to climate change, particularly during wildfire outbreaks (Ward *et al.,* 2016). Moreover, climatic conditions and topography have been found to influence SOC stock distribution (Odebiri *et al.,* 2020b; Hoffmann *et al.,* 2014). As such, the higher-altitude and wetter conditions of the sub-escarpment facilitates greater plant litter accumulation and increased SOC. Furthermore, the Northern KwaZulu-Natal Shrubland and Zululand Misbelt Grasslands, with finer clay soils, are also located within the sub-escarpment (Mucina and Rutherford 2006). Studies by Schwanghart and Jarmer (2011) and Singh *et al.,* (2011), have demonstrated a strong correlation amongst SOC and clay content. Soils with higher concentrations of clay have been observed to be more nutrient dense, thus affecting aboveground biomass production. The tight bondage between clay particles also protect SOC from microbial decomposition (Yao *et al.,* 2010). Consequently, the large pool of SOC located beneath South African grasslands can be attributed to a combination of favourable climate, topography and soil characteristics along the eastern coastal interior. Nevertheless, South African grasslands are prone to degradation, with only 2 % formally protected, which has significantly impacted SOC storage (Little *et al.,* 2015). For instance, Peri (2011) reported that heavy grazing can reduce carbon stock in grasslands by as much as 80 Mg C ha^{-1}. Thus, to safeguard SOC stocks within grasslands, adequate spatial planning and management frameworks supported by robust scientific data are necessary. The Durban Metropolitan Open Space System (DMOSS) framework is an example of a successful spatial instrument that has been used to ensure biodiversity needs are integrated into developmental planning (Roberts *et al.,* 2012). This tool is regularly updated through a joint partnership with the University of KwaZulu-Natal, which provides scientific research to inform and supplement environmental monitoring and management protocols within the municipality (Boon *et al.,* 2016).

Apart from Grasslands, tropical Savanna ecosystems account for more than 30% of South Africa's total SOC inventories (Table 6.4). The distinct open tree layer and herbaceous grassy substrate of savannas provide favourable conditions for SOC sequestration, as high quality litterfall and dead matter enable the faster metabolism of the resultant soil organic matter (De Deyn *et al.,* 2008). Meanwhile, the deep rooting zone of savanna vegetation protects against carbon losses through the topsoil and capable of sequestrating up to 110 t C/ha of SOC (Jackson *et al.,* 2000; Ciais *et al.,* 2011). In addition, the high clay content and swelling properties of the *latosols* within savannas promote SOC retention by providing physical and chemical mechanisms that shield SOC from microbial decomposition (Mujinya *et al.,* 2013; Shelukindo *et al.,* 2014). As with grasslands, the favourable topographic and climatic conditions along the eastern coastal interior enables higher concentrations of SOC stocks. Savannas, however, are threatened by land-use disturbance, with an estimated 1% lost to intensive grazing and cultivation per year (Gonzalez-Roglich *et al.,* 2015). Therefore, with the amount of carbon lost through degradation rapidly approaching tropical deforestation levels, appropriate carbon management strategies that incorporate grazing-land and fire management protocols are paramount. An example of this is the Kenyan Agricultural Carbon Project (Atela 2012). The project encourages sustainable and climate-friendly farming techniques that support soil carbon sequestration; and within three years of its inception, the project has increased agricultural productivity by 15%, whilst facilitating a reduction of 24,788 metric tons of carbon dioxide. Although the volume of land occupied by different biomes has been shown to influence SOC stock inventories, the relative sequestration capacity of these biomes varies greatly according to ecosystem dynamics (Laganiere *et al.*, 2013).

Despite their relative size, the Forest (44.12 t/h), the Indian Ocean Coastal Belt (43.05 t/h), and the Albany Thicket (38.76 t/h) biomes had the highest average SOC sequestration rates (Table 6.4). The Forest biome is generally comprised of multi-layer evergreen stands with high canopy coverage and a longer rooting residency — which can increase SOC sequestration through continuous litter fall (Muchena 2017). In forests, plant-carbon assimilation rates are high throughout the year due to an efficient use of soil nutrients and solar radiation that encourages continuous growth (De Deyn *et al.*, 2008). This high-rate of plant-carbon assimilation results in an elevated SOC sequestration capacity — with the majority of carbon stored within aboveground structures as opposed to the soil (Kesselmeier *et al.*, 2002). Additionally, an annual precipitation of more than 725 mm during the summer elevates the rate of carbon cycling and plant growth (Mucina and Rutherford, 2006). Although forests experience relatively low levels of wild-fires due to high humidity, a large percentage of soil organic matter and SOC are lost through deforestation and intentional burning (Low and Rebelo, 1998). For instance, according to the Global Forest Watch, about 30.3 Kha of South Africa's natural forest has been lost since 2010. This is the equivalent of approximately 11.3 Mt in additional carbon emissions (https://www.globalforestwatch.org/). Furthermore, an abundance of invasive lianas (*Annonacea*) plants, which create intricate webs along the understory of forest canopies, has the potential to limit carbon stocks by up to 50 % (Bousfield et al., 2020). Nevertheless, this is overshadowed by the need for housing and agricultural land which has caused significant deforestation and incalculable levels of SOC loss (Ekblad and Bastviken, 2019). To improve Forest SOC sequestration capacity and safeguard ecosystem services, well-managed indigenous reforestation projects are necessary. Literature (Lal, 2009, Olsson and Ardö, 2002, Don *et al.*, 2011; Venter *et al.*, 2021) has shown that the conversion of cultivated land to secondary forests or plantations may increase SOC stocks and facilitate climate change regulation. An example can be seen with the Buffelsdraai reforestation programme in Durban South Africa, where previously cultivated land has been reforested with indigenous species to facilitate carbon sequestration and boost overall ecosystem health (Mngadi *et al.*, 2022). Further evidence of the success of ecosystem-based mitigation activities can be found within programmes such as REDD+ (Dickson and Kapos, 2012).

Like Forests, the Indian Ocean Coastal Belt (43.05 t/h) and the Albany Thicket (38.76 t/h) biomes had high average SOC sequestration rates (Table 6.4). The Indian Ocean Coastal Belt consists of open to dense savanna vegetation, intermixed with vast areas of forest and sourveld grassland (Mucina and Rutherford, 2006). Its location along the coastal interior (with high mean annual rainfall and temperatures) allows for an accumulation of soil moisture. Elevated soil moisture concentrations foster anaerobic conditions, which accelerates SOC accumulation (Shelukindo *et al.,* 2014). The Albany Thicket biome is dominated by Spekboom (*Portulacaria afra*), a drought tolerant plant that binds soil together to minimize soil erosion (Mills *et al.,* 2015). It's rapid growth rate and high litter production encourages fast carbon sequestration potential of approximately 168 t C ha^{-1} from the top 50 cm of soil (Mills and Fey, 2004). However, intensive goat farming has led to significant thicket degradation, with between 40 to 71 t C ha^{-1} lost from the top 10 cm of soil (Mills and Fey, 2004).

Although both climatic and topographic variables are fundamental to SOC stock variability amongst biomes, the greatest threat to SOC accumulation and storage is land-use disturbance (Ramesh *et al.,* 2019; Were *et al.,* 2015). Degradation of natural biomes through urbanization and agricultural practices have significantly depleted SOC stocks (Schulze and Scuttle, 2020). It is estimated that approximately 200 Gt of carbon is released into the atmosphere as a direct result of land-use change and ecosystem conversion (Diversity, 2009). Lal (2008) notes that the transformation of ecosystems to agricultural land drastically depletes SOC, with agricultural soils retaining less than 50% of the SOC of undisrupted soils. In South Africa, Du Toit *et al.,* (1994) and Du Toit and Du Preez (1995), established that decreases in SOC from cultivated soil were highest in regions that were warm and dry, such as within the desert biome. Here, soil temperatures would impact the microbial mineralization of carbon. Likewise, Mills and Fey (2004), noted that vegetation removal drastically reduced SOC across most biomes, with SOC declining from 28 t ha^{-1} to 15 t ha^{-1} within renosterveld vegetation, from 7 t ha^{-1} to 5 t ha^{-1} in the Karoo, and from 54 t ha^{-1} to 27 t ha^{-1} in grasslands. With most of South African biomes already significantly altered by land-use change (Du Preez *et al.,* 2011, Ellis *et al.,* 2010), efforts should be tailored towards maintaining existing biomes and restoring degraded areas. Consequently, preventing ecosystem conversion and combating land degradation are practices that can shield current carbon stock inventories (Goldstein *et al.,* 2020). Besides, the restoration of certain degraded biomes has the potential to drastically bolster existing SOC stocks and improve SOC sequestration rates (Lal 2015). For instance, the use of Spekboom plant to restore degraded Thicket regions has the potential to sequester an additional 40 t C ha^{-1} (Mills and Fey, 2004). Similarly, invasive plant management has been gaining traction as an unconventional management technique for carbon sequestration (Peh *et al.,* 2015, Turpie *et al.,* 2008). Although some studies have suggested that invasive species could potentially increase carbon storage, they can also drastically alter ecosystems and ecosystem function (Liao *et al.,* 2008). An investigation by Koteen *et al.,* (2011) on invasive grasses within Californian grasslands has shown that invasive annual plants that replace drought-tolerant perennial grasses may reduce SOC storage by up to 40 t C ha^{-1}. Thus, further investigation on the long-term effects of invasion alien species on SOC is necessary within tropical biomes.

Nonetheless, the ability of specific biomes to increase carbon sequestration pools within soils is dependent upon an amalgamation of climate, topography and land-use disturbance (Lal, 2008). Thus, as populations grow and land-use conflicts intensify, evidence-based research in conjunction with comprehensive land-use planning frameworks are needed to provide an appropriate balance between carbon sequestration, agriculture, urbanization and biodiversity. At a project level, biodiversity assessment schemes in conjunction with decision support tools (such as UNEP, CBD, and LifeWeb carbon calculator) can be used to promote best practices (Bagstad *et al.,* 2013), however at a national level, these cannot replace the need for large-scale data on high-carbon density biomes.

6.4.2. Performance of the SOC CAE-DNN model

This study used a hybrid DL framework (Concrete Autoencoder-Deep neural network (CAE-DNN)) for variable selection and regression to create a spatial SOC variability map based on remote sensing-topo-climate data. The results were compared to those of alternate approaches that included the Boruta-DNN and a standard DNN with no variable selection method applied. The CAE-DNN (RMSE = 7.91 t/h) and Boruta-DNN (RMSE = 9.48 t/h) models outperformed the standard DNN (RMSE = 10.71 t/h). This demonstrates the importance of dimension reduction within model performance. More importantly, the CAE-DNN with the least selected variables (*n* = 26) performed better than the Boruta-DNN (*n* = 31) method. This further illustrated that DL variable selection methods generally outperform conventional variable selection strategies as supported by existing literature (Abid *et al.,* 2019; Ramjee *et al.,* 2019). The CAE employs concrete random variables and the re-parametrization method to allow gradients to flow through a layer that stochastically picks discrete input. This stochasticity enables it to swiftly explore and converge on a subset of input characteristics of a certain size that minimizes a specific loss (Abid *et al.,* 2019). The CAE is likewise straightforward to use, requiring only a minor modification of a typical autoencoder with a similar runtime — depending on hardware capacity (Abid *et al.,* 2019). The performance of the hybrid CAE-DNN framework is further evidenced by the results obtained by the study (R^2 = 69.9). This is a notable improvement to the national scale SOC models previously developed by Odebiri *et al.,* (2022) (R^2 = 67.3), Venter *et al.* (2021) (R^2 = 65.9) and Schultze & Schutte (2020) (R^2 = 20.3) using ordinary DNN models, random forest algorithm and field-level SOC median calculation strategy, respectively.

Subsequently, the SHAP technique (see section 6.2.5) was included within the overall hybrid DL framework to assess variable importance and evaluate the most relevant explanatory factors that determine SOC distribution within the region. Rainfall, NDVI, Band 8 (Red - 665 nm), Elevation, Temperature, EVI, Band 11 (red-edge - 708.75 nm), Slope, RVI, and TWI were among the top 10 variables identified. Rainfall and TWI are both indicators of water availability and soil moisture within different biomes. Rainfall impacts soil moisture, hydrological processes, vegetation density and decomposition, all of which help to sequester SOC (Chen *et al.,* 2015). TWI affects the variability of soil moisture along slopes, hence locations with a greater TWI (soil moisture) have a higher SOC density than those with a lower TWI (Odebiri *et al.,* 2020b; Li *et al.,* 2018). SOC was similarly influenced by temperature, as warmer areas accelerate SOC mineralization compared to areas with lower temperature (Wang *et al.,* 2013). Additionally, greater temperatures indicate higher atmospheric water vapour, which increases the likelihood of rainfall (Cong and Brady 2012). The NDVI, EVI, RVI as well as Band 8 (Red - 665 nm) and Band 11 (red-edge - 708.75 nm) are all the derivatives of Sentinel-3 data used in this study, which are considered sensitive to vegetation properties including biomass and chlorophyll content (Odebiri *et al.,* 2021a; Zhang *et al.,* 2019). Literature has established that SOC responds to similar physical and environmental signals as vegetation, thus can be used as a surrogate or proxy to predict its distribution (Nabiollahi *et al.,* 2019; Kumar *et al.,* 2016). Both Elevation and Slope relate to the altitude and steepness of the ground. We found that places with steep slopes have less occurrences of SOC than locations with relatively gentle slopes, and that areas with greater altitude (> 1500m) have lower SOC concentrations compared to low lying areas (< 1000m) (Li *et al.,* 2018; Weiss *et al.,* 2017). Low-lying regions enhance plant growth through favourable soil conditions including the transportation of nutrient-rich topsoil from higher to lower lands. Furthermore, most low-lying regions have greater soil moisture content, nutrients, and deeper than higher areas characterized by harsh environmental conditions that inhibit vegetation growth due to a lack of soil micro-organisms (Hijmans and Graham 2006). The impact of elevation and slope on microclimate could possibly be a factor in SOC concentration, since altitude critically affects temperature levels, wind flows, and soil moisture, all of which affect vegetation distribution and by extension SOC stocks (Odebiri *et al.,* 2020b).

To maintain a balance between carbon sequestration capacity, biodiversity, and the provision of ecosystem services, accurate national-scale SOC assessments are required periodically. Consequently, this study has demonstrated the ability of the Concrete Autoencoder-Deep Neural Networks framework and remotely-sensed-topo-climatic datasets in providing cost-effective and reliable SOC models for recurrent carbon assessments. Lastly, as strides within modern satellite sensor technology and deep learning architecture are made, robust semi-autonomous tools for understanding the spatial dynamics of SOC are constantly evolving.

6.5. Conclusion

In this study, different remotely sensed topo-climate variables were used to undertake a national scale SOC modelling across different South African biomes. For variable selection and regression, a combination of DL frameworks — Concrete Autoencoder-Deep neural network (CAE-DNN) performed better than Boruta-DNN and a standard DNN without variable selection. The estimated SOC stocks within each biome, as well as their potential sequestration rates in proportion to area coverage were established. This is crucial information for national policy formulation, rehabilitation, restoration, and intervention initiatives aimed at mitigating climate change by increasing SOC assimilation. We also implemented the SHAP value approach to resolve DL models' interpretability problems, with Rainfall, NDVI, Band 8, Elevation, Temperature, EVI, Band 11, Slope, RVI, and TWI among the top 10 most important variables. Taking into account the spatial uncertainty of the data utilized in this study, including multiple-scales of variation, diverse sampling sources, and different collecting times, our model performed quite well. This necessitates that efforts by government and research-funded bodies be geared towards standardized soil inventory systems in which high-quality data specific to diverse biomes are collected and made available for use in SOC modelling. However, when adopting DL frameworks for simulation, cost and sample size should be considered. This is because DL applications demands high computing capacity and big data to perform optimally, specifically for landscape scale mapping. Future research could benefit from examining the ability of greater spatial resolution sensors, as well as the fusion or pan sharpening of many sensors (e.g. Sentinel 1 and 2), to increase accuracy when utilized with deep learning (DL) models. Finally, SOC future projection studies could also be conducted to help effectively evaluate the efficacy of current management systems and policies.

6.6. Summary

This chapter examined the capabilities of two DL frameworks (Concrete Autoencoder-Deep neural network (CAE-DNN) for variable selection and regression) in conjunction with remotely sensed topo-climate variables to model SOC across South Africa's unique biomes. The CAE-DNN model combined with climatic and terrain metrics was found to significantly outperform both the Boruta-DNN and the standard DNN model without variable selection. SOC stock was shown to correlate to overall biome coverage, with the Grassland and Savanna biomes contributing the most to South Africa's overall SOC pool. Nonetheless, despite their lower footprint, the Forests and the Indian Ocean Coastal Belt biomes demonstrated the highest SOC sequestration capacity. These findings outline the importance of large bio-climatic zones in regional SOC storage and highlights the need to direct managerial efforts towards promoting SOC sequestration across different biomes to extenuate the effects of climate change. However, the dynamic nature of climate change and land-use transformation presents a unique threat to current SOC stocks and management techniques. Consequently, to effectively safeguard regional SOC stocks, policymakers require large-scale projections that accurately model both current and future SOC pools. In this regard, the next chapter investigates the development of large-scale projections for both existing and future SOC pools in South Africa. This will assist policymakers in assessing the long-term viability of South Africa's existing SOC management techniques and land-use planning frameworks.

Chapter Seven:

Evaluation of projected soil organic carbon stocks under future climate and land cover changes in South Africa using a deep learning approach

This chapter is based on;

Odebiri, O., Mutanga, O., Odindi, J., & Naicker, R. (2022). Evaluation of projected soil organic carbon stocks under future climate and land cover changes in South Africa using a deep learning approach. *Journal of Environmental Management*, Under Review, Manuscript ID: **JEMA-S-22-05916**

Abstract:

Environmental degradation and carbon emissions have become a major global concern. This has forced policymakers to consider strategic and long-term contingencies to increase carbon sequestration capacity and mitigate the effects of climate change. Soil organic carbon (SOC) provides a reliable long-lasting mechanism to ameliorate climate change and regulate carbon fluxes. However, unanticipated rates of climate change coupled with the dynamic nature of land-use transformation threatens current mitigation approaches and can jeopardise carbon stock assimilation. To effectively manage and protect SOC stocks, large-scale projections that accurately model both current and future SOC pools are necessary. Hence, this study modelled the effects of simulated climate and land-cover change on SOC inventories across South Africa up to the year 2050. A digital soil mapping strategy in concert with a deep neural network (DNN) was used to model current SOC stocks distribution. Subsequently, WorldClim general circulation models and a space-for-time substitution (SFTS) method were used to derive future SOC stocks under four shared socio-economic emission pathways. Depending on emission rates, results showed a reduction in SOC inventories, with overall SOC stocks declining from 5.64 Pg to between 4.97 and 5.38 Pg by 2050. Meanwhile, forests, which account for approximately 1.2 Pg of total SOC in South Africa, were found to have lost more than 1% of their total coverage by 2050. These findings provide a glimpse into the state of South Africa's current and future SOC stock inventories and the influence of climate and land-use change. These findings are valuable to among others policymakers, land use managers and climate change experts in assessing the long-term feasibility of South Africa's existing SOC management protocols and land-use planning agenda. However, to adequately protect future SOC stocks, current land-use planning frameworks need to be re-adjusted to prioritize pressing environmental concerns.

Keywords: Soil organic carbon; Climate; Land cover; Topography; Deep learning; Management

7.1. Introduction

Higher rates of carbon emissions and rampant environmental degradation have left many regions vulnerable to severe climatic disasters (Dietz, 2020; Naidoo & Fisher, 2020). Over the past 50 years for instance, climatic disasters have resulted in over 2 million deaths and USD 3.64 trillion in damage (WMO, 2021), with a further 4 % of global annual GDP projected to be lost within the next 30 years (Newell *et al.,* 2021). As exposure to floods, wildfires, and other extreme weather-related events becomes more prevalent worldwide (Arletti *et al.,* 2021), governments and policymakers are forced to fast-track reliable, long-lasting strategies to mitigate the effects of climate change and minimize both human and financial losses. In this regard, soil organic carbon (SOC) sequestration has emerged as one of the most important carbon capture and storage processes for mitigating climate change, with approximately 1500 Petagrams (Pg) of carbon stored in the top 1-meter of soil globally (Padarian *et al.,* 2021, Li *et al.,* 2021). SOC constitutes more than 61% of total soil carbon and plays an essential role in regulating global carbon fluxes (Zhao *et al.,* 2021; Odebiri *et al.,* 2022a). Nevertheless, due to its central role within the global carbon exchange, slight changes in SOC stocks can have a significant impact on climate and ecosystem stability (Balesdent *et al.,* 2018; Wu *et al.,* 2018). As a result of this dynamic interchange between the terrestrial carbon cycle and climate change, there is a need to constantly evaluate SOC inventories and management practices (Ren *et al.,* 2020; Odebiri *et al.,* 2021b). However, adequately incorporating SOC as a long-term climate change mitigation strategy requires a comprehensive large-scale understanding of both present and future SOC stocks, as well as the influence of environmental and anthropogenic factors (Keskin *et al.,* 2019).

The storage and sequestration potential of SOC is often dependent on topographic, climatic, and land-use factors, which control the overall viability and variability of SOC stocks (Zhao *et al.,* 2021; Yigini & Panagos 2016). Literature has revealed that topo-climatic factors (i.e. rainfall, temperature, slope, elevation, and relief) can significantly affect organic compound accumulation and decomposition — thereby regulating soil microbial activity and SOC dispersal (Schindlbacher *et al.,* 2012; Reyna-Bowen *et al.,* 2020; Guan *et al.,* 2018; Conant *et al.,* 2011; Li *et al.,* 2018; Fissore *et al.,* 2017; Wang *et al.,* 2012). On the other hand, land-use change factors (in the form of urbanization, intensive agriculture, and increased environmental degradation) has a more profound influence over SOC and atmospheric carbon emissions (Lamichhane *et al.,* 2019; Swanepoel *et al.,* 2016; Sainepo *et al.,* 2018; Odebiri *et al.,* 2021b; IPCC *et al.,* 2021). For instance, Amanuel *et al.,* (2018), noted that changes in land use accounts for between 12 and 20 percent of human-induced carbon emissions while Sanderman *et al.,* (2017) noted that during the last decamillennium, over 30 Pg of carbon has been lost to intensive agronomic practises. Although, SOC variability, as a direct result of land-use and climate change, has been acknowledged as a key issue in climate change mitigation and environmental sustainability (Jiao *et al.,* 2020; Ou *et al.,* 2017), the complex and interconnected nature of these threats on soil and carbon security requires a deeper temporal contextualisation at a regional scale (Keskin *et al.,* 2019).

South Africa, ranked as the 13th largest greenhouse gas contributor worldwide (Jäättelä *et al.,* 2017), has been subjected to considerable SOC loss due to land cover transformation (Venter *et al.,* 2017). For example, Griscom *et al.,* (2017) noted that around 2 Pg of South Africa's carbon has been lost to land-use change since 10,000 BCE, while Du Preez *et al.,* (2011) reports that South Africa’s soil carbon reserves have diminished by 45 percent due to continuous agriculture. Recent investigations by Schulze and Scuttle, (2020) and Venter *et al.,* (2021) have further outlined the detrimental impact of land use transformation on SOC stocks in South Africa. Schulze and Scuttle, (2020) for instance established that almost half of all SOC stored within natural vegetation have been lost to intensive agriculture. Although these studies have provided insight into the variability of SOC stocks across South Africa, the rapidly evolving nature of climatic and land-use change within South Africa has prompted researchers to now contemplate the influence of future climate and land-use change scenarios on current SOC stock inventories. Such knowledge is needed to adjust existing policies and SOC management practices to cope with accelerated climate and land use change transitions within South Africa (Wiesmeier *et al.,* 2016). Yet, to the best of our knowledge, no known study has attempted to examine the influence of potential future climate and land use changes scenarios on SOC variability across South Africa. However, making accurate predictions in highly dynamic and complex environments, such as soils, can be challenging over large geographical extents (Padarian *et al.*, 2020; Odebiri *et al.*, 2020a), particularly as soil datasets are frequently out of date, have limited coverage, and are mostly fragmented (Zhao *et al.*, 2021).

In recent years, digital soil mapping (DSM) has emerged as a science connecting proximal soil observations with quantitative methodologies, such as the use of novel earth observation satellites with high spatial and spectral coverage (i.e. big data) in conjunction with various geostatistical and machine learning (ML) models (Georgiou *et al.*, 2021; Guo *et al.*, 2020; Xu *et al.,* 2019). As a result, DSM has gained traction as a credible means of obtaining soil data at various levels of detail (Wieder *et al.,* 2018; Muchena 2017: Kumar *et al.,* 2016). This has facilitated the simulation of SOC across multiple temporal scales using data-driven empirical models (DDEMs) (Wang *et al.,* 2021; Wang *et al.,* 2019). Specifically, Space-For-Time Substitution (SFTS), which forms the basis of DDEMs within temporal modelling dynamics, provides comprehensive covariate substitution capabilities (Li *et al.,* 2021; Huang *et al.,* 2019; Gray & Bishop 2016). This has allowed for future SOC models to be predicted through a mathematical substitution of present covariates with future covariates that have been derived from DDEMs incorporating SOC measurements and spatially variable environmental factors (Wang *et al.,* 2021; Wang *et al.,* 2019). However, the ability of these DDEMs to effectively represent SOC distribution successfully is predicated upon dataset preparation techniques and the algorithm used (Odebiri *et al.,* 2022a; Wadoux *et al.,* 2019). Furthermore, to minimize uncertainties between long-term field measurements and SOC stock models, Sulman *et al.,* (2018) and Wieder *et al.,* (2018), have recommended the development and implementation of benchmarking models to improve model accuracy.

Deep learning (DL) models, which utilize non-linear representational-learning functions to extract information from lower to higher layers, has been lauded as robust analytical tools for improving SOC mapping (Odebiri *et al.*, 2021a; Ma *et al.*, 2019; Zhu *et al.*, 2019). DL frameworks, unlike other geostatistical and traditional machine learning (ML) algorithms, can use feature representations learnt solely from input data to considerably improve mapping capabilities (Odebiri *et al.*, 2021b). Additionally, the complicated nonlinear structure of SOC and its relationship with the environment can often present a high degree of variability, which can overwhelm traditional machine learning methods, especially at large spatial scales (Ma *et al.,* 2019; Padarian *et al.,* 2019; Kumar *et al.,* 2016). In contrast, DL models can improve learning procedures by identifying relevant invariant and abstract qualities from datasets (Minh *et al.,* 2018). The inclusion of numerous hyper-parameters in DL models also allow users to fine-tune learning techniques to improve accuracy (Odebiri *et al.,* 2021a). To this end, this study used multi-source data and a DL algorithm to project SOC stock distribution under future climatic and land-use change scenarios to the year 2050. Thereafter, we evaluated and compared the variability of current (2021) and future (2050) SOC stocks across key land uses in South Africa.

7.2. Methods and Materials

7.2.1. Soil data

The International Soil Reference Information Centre (ISRIC) and the Department of Agricultural Earth and Environmental Sciences (SAEES) at the University of KwaZulu-Natal in South Africa provided the soil data for this study. ISRIC is a non-profit science organization whose purpose is to deliver high-quality information on diverse soil properties (including SOC) through international collaboration (Batjes *et al.,* 2020). The present ISRIC soil database (https://www.isric.org/), which was last updated in 2020, has over 150,000 sample points from 173 countries (Batjes *et al.,* 2020). The soil profile from the majority of these sites may differ in terms of location and acquisition time. To circumvent this and other potential discrepancies in SOC concentration determination methodologies, a standardized framework, publicly available at https://www.isric.org/explore/wosis/accessing-wosis-derived-datasets, was developed by ISRIC to enable consistent soil profile data entry (Venter *et al.*, 2021; Hengl *et al.,* 2017). From the ISRIC and SAEES databases, a total of 1936 soil sample points (collected at a 30 cm depth) were collated and input as target variables. Thereafter, following methods used by Pearson *et al.*, (2007), the SOC stock at each point was calculated through the following formula:

$$SOC\ accumulation\ (t/h) = SOC\ concemtration\ x\ volume\ density\ x\ soil\ depth \qquad (1)$$

7.2.2. Simulation data acquisition

7.2.2.1. Sentinel 3

The study initially developed a current SOC estimation model (2021) to understand the variability of SOC across South Africa. This then provided a basis to form the future SOC projection model (2050), whilst simultaneously allowing the study to evaluate the potential influence of topo-climatic and land-use change factors on South Africa's SOC stocks over the next few decades. In accordance with this, data from the European Space Agency's (ESA) Sentinel-3 Ocean and Land Colour Instrument (OLCI), with under 10% cloud coverage, was obtained between March and April 2021 from the ESA portal (i.e. https://scihub.copernicus.eu/). Sentinel-3, which provides 21 unique spectral bands (400 nm to 1020 nm), a 300 m spatial resolution, and a 2 days' revisit time enables continuous, reliable, and fast regional SOC mapping (Vaudour *et al.*, 2019; Li and Roy, 2017). The Sentinel-3 imagery was geometrically, radiometrically, and atmospherically corrected using the Sentinel Application Platform (SNAP v. 6.0). The pre-processed images were mosaicked and clipped to the South African border. Studies by Wang *et al.,* 2021, Odebiri *et al.,* 2020a, and Guo *et al.*, 2020, have demonstrated the value of vegetation indices in explaining SOC fluctuation and dispersion. Consequently, using information derived from Odebiri *et al.*, 2022a, the most influential Sentinel-3 spectral bands (i.e. 6, 7, 8, 11, 13, 17, 18, and 19) and vegetation indices ((including the Normalised Difference Vegetation Index (NDVI), the Enhanced Vegetation

Index (EVI), the Soil Adjusted Vegetation Index (SAVI), the Difference Vegetation Index (DVI) and the Ratio Vegetation Index (RVI)) were identified (n = 13) and used to construct the current SOC estimation model (2021).

7.2.2.2. Topo-climate metrics and climate scenarios

Topographic and climatic factors, which can widely vary, have been documented as important determinants of SOC variability (Li *et al.,* 2018; Fissore *et al.,* 2017; Wang *et al.,* 2012). However, temporally, studies by Liu *et al.,* (2021), Huang *et al.,* (2019), Gray & Bishop (2016), and Yigini & Panagos (2016) have shown topographical variables to remain relatively stable over time. Hence, using information obtained from previous investigations by Odebiri *et al.*, (2020b), the study identified eight influential terrain metrics that characterise the study area (i.e. Slope, Elevation, Aspect, Topographic Wetness Index (TWI), General curvature, Catchment Area, Profile curvature and Direct insolation). These variables were derived from a Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) using the SAGA GIS (2.3.2) and ArcGIS Pro 2.8 software and incorporated into both the current (2021) and future (2050) SOC models.

Both current and future climate representative data (mean annual rainfall and temperature) for South Africa were obtained from the one square kilometre (1km^2) 30 arc-seconds spatial resolution of the WorldClim database (http://www.worlclim.org/). For future SOC projections, the model was run using an average of five WorldClim General Circulation Models (GCM) (i.e. CNRM-CM61-1, CanESM5, GFDL-ESM4, ACCESS-ESM1-5, and INM-CM5-0) under four Shared Socio-economic Pathways (SSPs): 126 (low emission pathway), 245 (low-to-moderate emission pathway), 370 (moderate-to-high emission pathway) and 585 (high emission pathway) (Liu *et al.*, 2021; Caddeo *et al.*, 2019; Gray & Bishop 2016).

7.2.2.3. Projected land cover

Changes in land cover has been documented by several studies to considerably influence SOC sequestration and storage capacity (Lamichhane *et al.*, 2019; Swanepoel *et al.*, 2016; Amanuel *et al.*, 2018). Subsequently, this study used a projection procedure outlined by Kamaraj and Rangarajan, (2022) to develop South African land use maps for the years 2021 and 2050. The projection process involved first obtaining past and present (i.e., 2014, 2018 and 2021) land cover maps from the South Africa National Land-Cover Map's (SANLC) online portal (https://egis.environment.gov.za/gis_data_downloads). These maps (with an average accuracy of >80%) were developed from robust satellite imagery over a multi-temporal time period.

These maps were then reclassified into eight major classes (Figure 7.1, Table 7.1) based on categorical descriptions derived from the South Africa National Land-Cover Map (SANLC, 2020). These categories were purposefully chosen due to their dynamic influence over SOC stocks. These classes included grassland, natural forest, commercial forest, crop land, shrubland, barren land, built-up vegetation and other (SANLC, 2020). Next, the SANLC land cover maps (SANLC 2014 and SANLC 2018), together with other spatial variables such as DEM, slope, aspect and distance from road and rivers were input into a deep neural network (DNN) workflow. Using the following hyper-parameters: iterations — 1000, neighbourhood — 1, hidden layer —10, momentum value — 0.06, learning rate — 0.001 (Kamaraj and Rangarajan, 2022; Das and Sarkar 2019; El-Tantawi *et al.,* 2019), a projected land cover map was derived for the year 2021 based on the evolving patterns identified from the preceding years (i.e. 2014 and 2018). This was then validated using the SANLC 2021 land cover map. Having established a high correlation (% of correctness — 84.28, kappa overall — 0.804) between the SANLC 2021 and the new projected 2021 land cover map, we conducted a new land cover projection to the year 2050. The projected 2050 land cover map was then added as an input variable to predict potential SOC for 2050. As a final step of data pre-processing, all the acquired and generated simulation data were resampled to a 300 meters' spatial resolution.

Table 7.1. Land use classes and their description derived from South Africa National Land-Cover Map (SANLC 2021)

Land use type	**Description**
Natural forest	Contiguous (indigenous) forest, contiguous low forest and thicket, dense forest and woodland, and open woodland
Commercial forest	Contiguous and dense plantation forest, open and sparse plantation forest, and temporary unplanted (clear-felled) plantation forest are all examples of plantation forest used for commercial purposes.
Grassland	Natural grassland with few trees and thinly forested grassland
Built-up vegetation	All sorts of vegetation including trees, bush, grass and bare within rural and urban setting
Cropland	Permanent crops, transitory crops, and vacant lands/old fields are all examples of cultivated permanent crops.
Barren land	Dry pans, degraded areas, terrestrial sand and dunes, coastal sand and dunes, bare riverbed material, and other bare surfaces
Shrubland	Low woodland communities with a canopy height of 0.2 to 2 meters in the wild
Other	Water bodies, buildings/mines structures without possibility of soil organic carbon

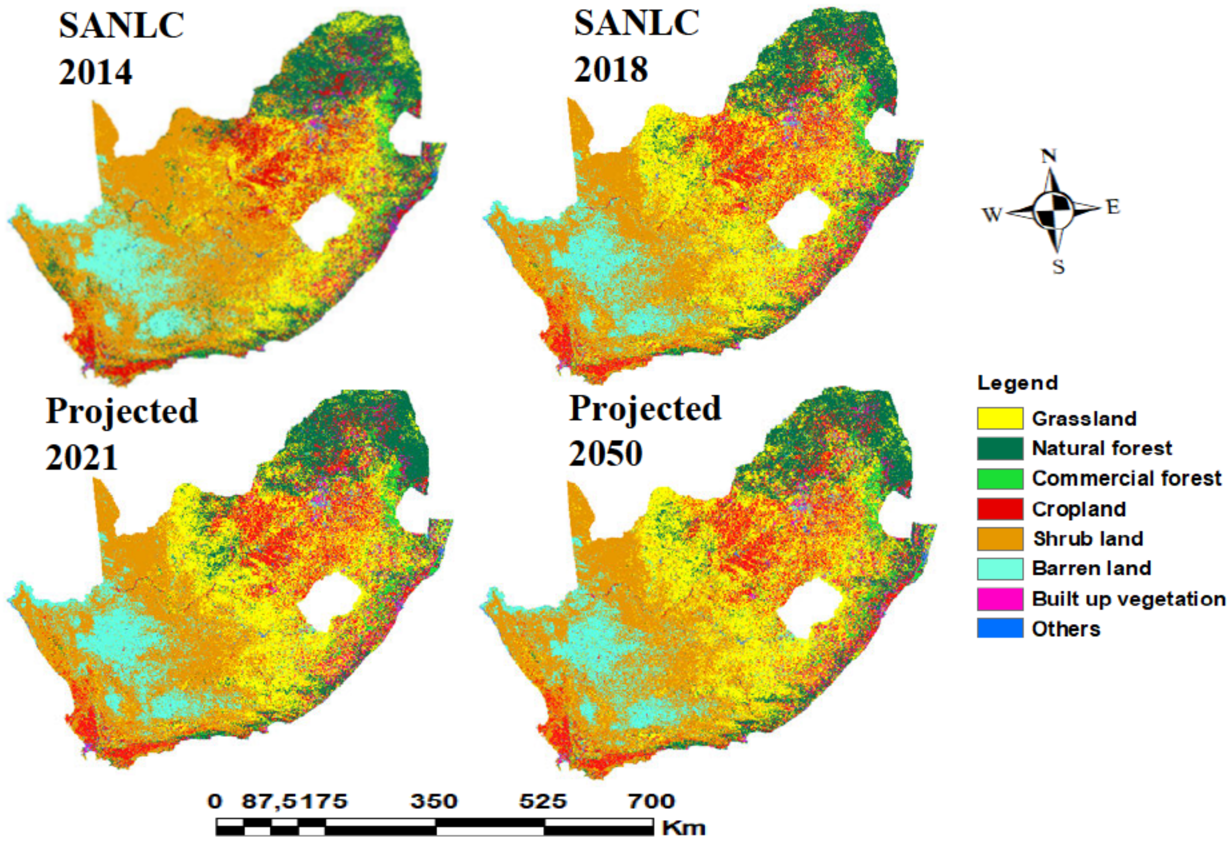


Figure 7.1. Projected South African land cover from the year 2014 to 2050, depicting the distribution of the eight key land uses (i.e. natural forest, commercial forest, grassland, built-up vegetation, cropland, shrubland, barren land and other)

7.2.3. Deep Neural Networks (DNN) and SOC model development

In recent years, the use of deep neural network (DNNs) architecture within SOC mapping activities has gained considerable attention (Emadi *et al.,* 2020; Pudełko & Chodak., 2020). In its most rudimentary configuration, a DNN model uses multilayer perceptron architecture (MLP) comprised of an input layer, many hidden layers, and an output layer (Taghizadeh-Mehrjardi *et al.,* 2020). This forward-feeding machine learning mechanism enables the connections between neurons within each of the different layers to fully connect with each successive layer, providing information regarding linear and non-linear interactions between the response variable and a set of predictor variables (Wang *et al.,* 2020).

For this study, a total of 24 input variables were used to model current SOC stocks across South Africa and its major land uses. These included eight Sentinel-3 spectral bands, five vegetation indices, ten topo-climate variables, and the projected South African land-use map (2021). This dataset was then divided into ten equal segments and sequentially utilized for training and testing purposes. The DNN was calibrated ten times to ensure that each data point was used for validation at least once. The "adam" optimizer, the ReLU activation function, 500 epochs, and six hidden layers were utilized for the best output after employing a random search hyper-parameter optimization with 10-fold cross-validation. For the development of the future SOC prediction model (2050), the projected WorldClim climatic scenarios, topography and the projected land cover map (2050) were substituted for the variables used within the previous DNN model. The DNN model is schematically represented within the base model phase workflow in Figure 7.2. See Odebiri *et al.*, (2022a), for more detailed explanation of the DNN model, including a mathematical representation and the hyper-parameter tuning method.

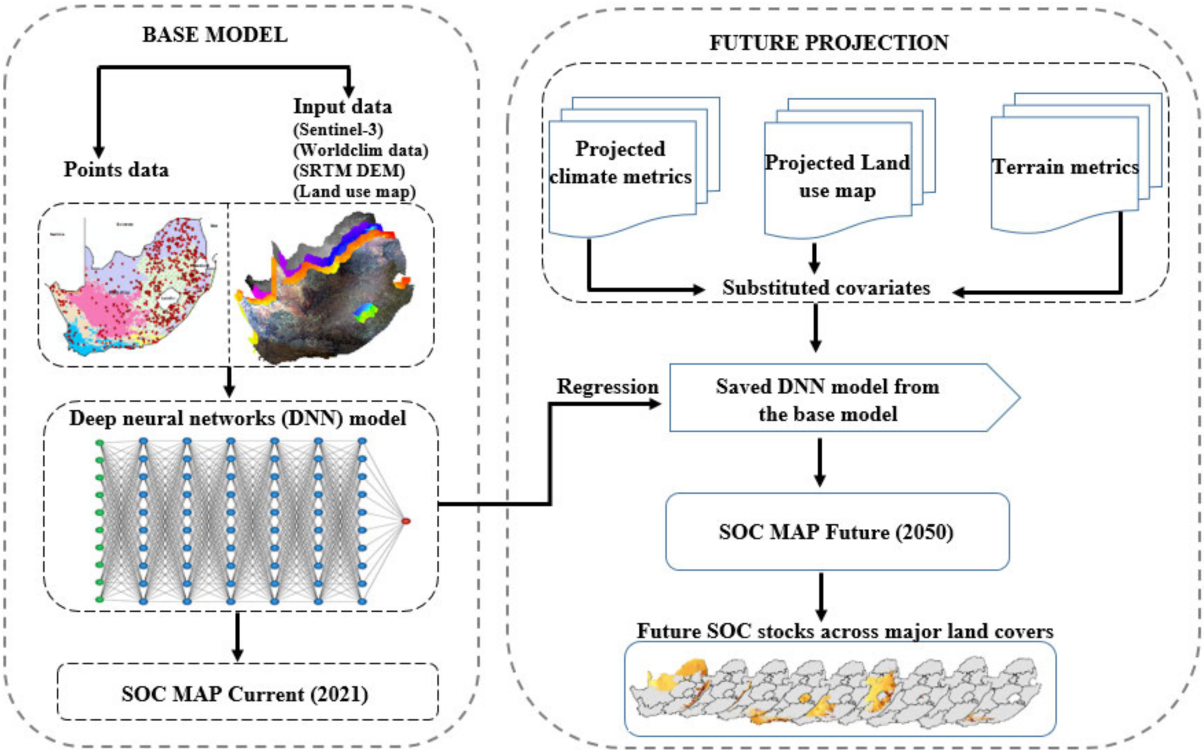


Figure 7.2. Soil organic carbon prediction and projection workflow

7.2.4. Model evaluation and uncertainty

The root mean square error (RMSE) and coefficient of determination (R^2) were used in this study to assess the fit and generalization of the DNN models. These evaluation metrics are expressed by the equations below:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (X_{O,i} - X_{P,i})^2}{n}} \qquad (2)$$

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} \quad (X_O - p)^2}{\sum_{i=1}^{n} \ (X_O - O')^2}\right] \tag{3}$$

Where n = number of observations, X_O and XP = measured and anticipated SOC values respectively. The averages of the measured and estimated SOC are represented by O' and P, while the variances of observed and predicted values are represented by $\boldsymbol{\sigma}_o$ and $\boldsymbol{\sigma}_p$. Thereafter, to ensure that the DNN model was free of sampling bias, a ten-fold cross-validation techniques was used. The data was separated into ten groups and assigned to calibration and validation data in a sequential order so that each variable was used at least once.

The SHapely Additive exPlanations (SHAP) methodology was then used to assess the importance of each predictor variable to the DNN model's overall performance (Pentoś 2016). The SHAP technique operates by assigning a specific magnitude of importance to each predictor variable with the DNN output model using its "DeepExplainer," feature. The SHAP method was chosen due to its robust global and local interpretability as compared to other techniques (for more on SHAP, see Lundberg and Lee, 2017).

To predict future (2050) SOC stock, the covariates of the present (2021) predictors within the DNN architecture was replaced with the projected covariates using the Space-For-Time Substitution (SFTS) assumption. As such, a total of eleven covariates including eight terrain metric, two climatic metrics and the projected land cover map for South Africa was used to simulate the SOC pattern for the year 2050 under four different Shared Socio-economic Pathways (SSPs).

7.3. Results

7.3.1. Model performance and current SOC stocks in South Africa

A total of 24 different spectral, topo-climatic and land-use variables were input into a deep neural network algorithm to model current (2021) SOC stock inventories in South Africa. After 10-fold cross-validation, the best SOC model produced an R^2 value of 0.7102 and an RMSE of 7.44 t/h (18.73 % of the mean) (Figure 7.3a). Thereafter, the SHAP procedure identified 10 significantly important variables (with a SHAP value ≥ 2) for determining SOC variability (Figure 7.3b). These included: Rainfall, Normalised Difference Vegetation Index (NDVI), Band 8 (665 nm), Elevation, Land cover, Temperature, Enhanced Vegetation Index (EVI), Slope, Band 11 (708.25 nm) and the Topographic Wetness Index (TWI). Figure 7.4 shows the spatial distribution of current (2021) SOC stocks across South Africa. This model highlights that South Africa currently has approximately 5.64 Petagrams (Pg) of total SOC stocks within its soils. SOC concentrations range from 0 to 147.24 t/h, with the majority of SOC densely situated along South Africa’s eastern coastline and adjacent interior.

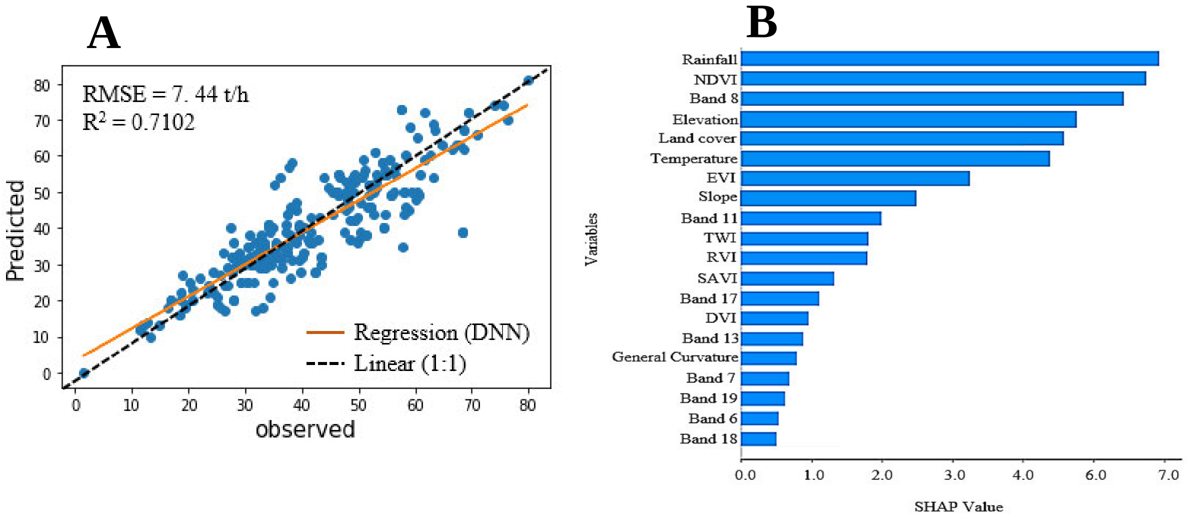


Figure 7.3. (A) correlation between the predicted and observed soil organic carbon (SOC); (B) importance ranking of predictors used for the simulation of SOC across South Africa

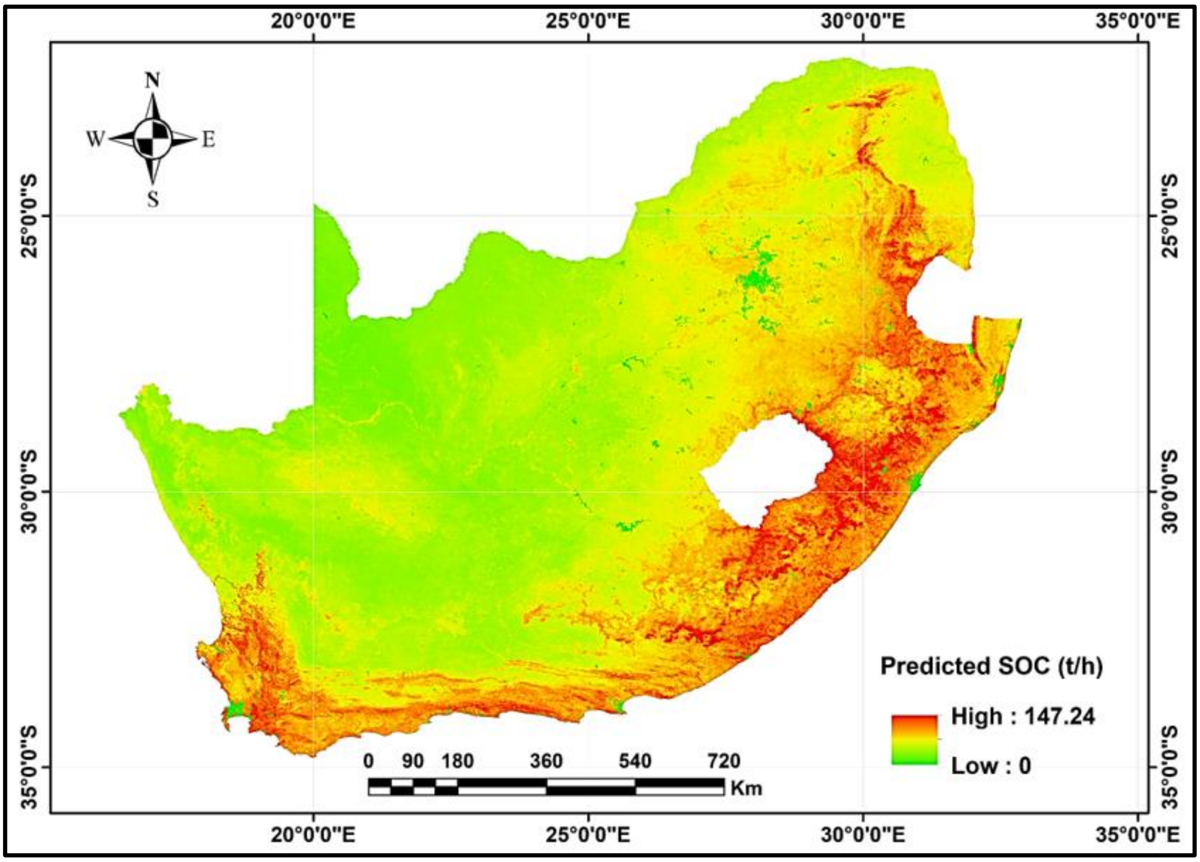


Figure 7.4. Map showing the current (2021) spatial distribution of SOC stocks across South Africa developed from the Deep Neural Network (DNN) model

7.3.2. Evaluation of future SOC stocks in South Africa

The estimated future SOC stocks (2050) differed amongst the four Shared Socio-economic Pathways, namely, SSP126, SSP245, SSP370, and SSP585. The lowest emission shared socio-economic pathway (SSP126) produced the best accuracy, with an R^2 value of 0.643 and an RMSE of 11.37 t/h (28 percent of the mean). SSP370 (medium to high emissions) was second ($R^2 = 0.622$, RMSE = 12.41 t/h — 31% of the mean), followed by SSP245 (medium emissions) with an $R^2 = 0.617$ and RMSE = 12.89 t/h — 32% of the mean, and lastly SSP585 (high emissions) ($R^2 = 0.598$, RMSE = 13.56 t/h — 34% of the mean). Figure 7.5 depicts the spatial pattern of SOC stocks across the different SSPs, with total SOC stocks estimated at 5.38 Pg (SSP126), 5.21 Pg (SSP245), 5.16 Pg (SSP370), and 4.97 Pg (SSP585), respectively. A clear downward trend in SOC stocks is evident amongst the different SSPs, with particularly low SOC inventories projected for the extreme climatic scenario (SSP585 – Figure 7.5d). A deeper examination of the maps reveals a major shift in SOC distribution and densities as the climate becomes more extreme. Specifically, climate change results in a notable reduction in SOC concentrations within the north-eastern, south-eastern, and south-western parts of South Africa. This downward trend of SOC stocks across the SSPs (Figure 7.5) correlates with a reduction in mean rainfall and an increase in temperature (Table 7.2).

Table 7.2. Mean rainfall and temperature for the current (2021) and future (2050) climate scenarios under four SSPs. The values in parenthesis connotes effective change between scenarios (with the negative values indicating a decrease).

Climate Scenarios	Scenario Extreme	Year of Scenario	Mean Rainfall (mm)	Mean Temperature (°C)
Current	—	2021	481.47	17.54
SSP126	Low	2050	474.26 (-7.21)	19.73 (2.19)
SSP245	Medium	2050	466.81(-14.66)	20.02 (2.48)
SSP370	Medium to High	2050	469.14 (-12.33)	19.91 (2.37)
SSP585	High	2050	432.29 (-49.18)	24.26 (6.72)

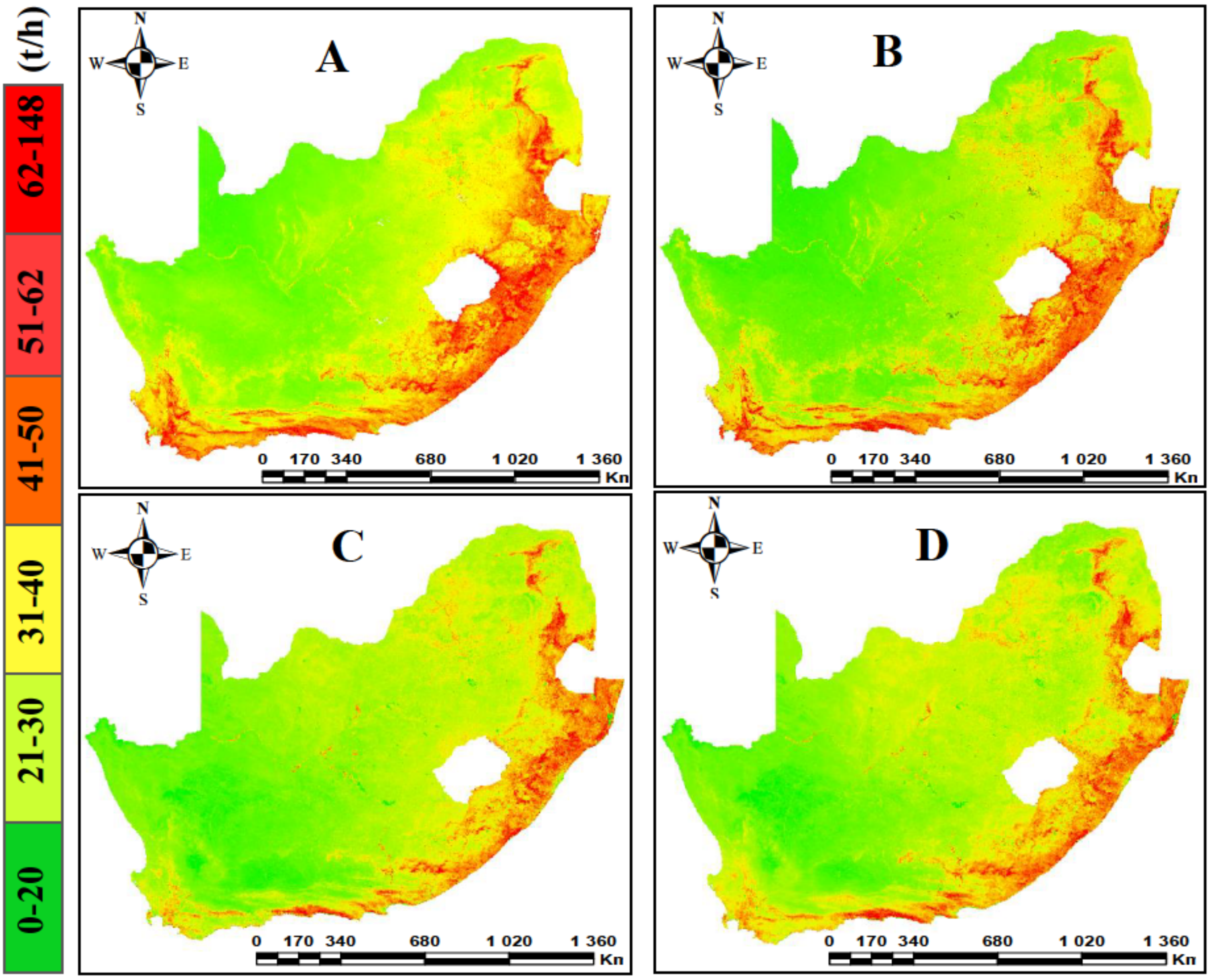


Figure 7.5. Spatial occurrence of SOC stocks under the four Shared Socio-economic Pathways (SSPs) namely; (A) SSP126, (B) SSP245, (C) SSP370 and (D) SSP585

7.3.3. Assessment of current and future SOC stocks across major South African land uses

Table 7.3 depicts the amount of land currently (2021) occupied by each individual land use and potential future coverage of each land use for the year 2050. The majority of South Africa's 1.2 million square kilometres of terrestrial landmass is currently occupied by grasslands (27.99 %) and shrublands (26.34%). In contrast, natural forest (~16%), cropland (~15%), and barren terrain (11%) individually occupy a relatively modest coverage. The smallest areas were built up vegetation and commercial plantations, which together accounted for approximately 4% of the total area (Table 7.3). A comparison between current (2021) and future (2050) land covers revealed substantial changes among the different classes. For instance, the area occupied by natural and commercial forests is projected to decline by 0.94 % (13198.04 km^2) and 0.12 % (1612.05 km^2), respectively. Similarly, built-up vegetated areas, and barren regions are expected to decrease by 0.33 percent and 0.57 percent (Table 7.3). However, over the next few decades, the area occupied by grasslands (0.76 %), croplands (0.61%), and shrublands (0.59%) are likely to increase.

Table 7.3. Projected land cover between 2021 and 2050. Changes are indicated in both square kilometres (Km^2) and percentage change(Δ); with a negative sign signifying a decrease in land area.

Class	2021 (Km^2)	2050 (Km^2)	Δ (Km^2)	2021 (%)	2050 (%)	Δ (%)
Grassland	332777.85	338013.29	5235.44	27.99	28.75	0.76
Natural Forest	188688.91	175490.87	-13198.04	15.87	14.93	-0.94
Commercial Forest	20694.85	19082.80	-1612.05	1.74	1.62	-0.12
Cropland	178098.61	183290.15	5191.54	14.98	15.59	0.61
Shrubland	313157.52	316551.05	3393.53	26.34	26.93	0.59
Barren land	127704.74	119585.16	-8119.58	10.74	10.17	-0.57
Built up vegetation	27793.72	23592.44	-4201.28	2.34	2.01	-0.33
Total	**1188916.20**	**1175605.76**		**100.00**	**100.00**	

Table 7.4 and 7.5 display the total SOC stocks (Pg) as well as the mean SOC density (t/h) for each land cover type across each shared socio-economic pathway (SSP) for the years 2021 and 2050. Grasslands (1.72 Pg) and Shrublands (1.12 Pg) had the largest SOC stocks amongst all of the land uses. These significant SOC pools are attributed to the large area occupied by these classes (Table 7.4). Other classes exhibited a similar trend, with total SOC stocks correlating to the percentage of land occupied (Table 7.4). For instance, natural forest and cropland, which occupy 15.87 % and 14.98 % of the total landmass, respectively (Table 7.3), contributed 1.04 Pg and 0.87 Pg to South Africa's total SOC pool (Table 7.4). In terms of SOC sequestration potential (illustrated by mean SOC density), natural (45.01 t/h) and commercial (46.95 t/h) forests sequester significantly more carbon in comparison to the other classes (Table 7.5). Meanwhile, built-up vegetated areas (38.32 t/h), cropland (34.55 t/h) and grassland (30.72 t/h) had moderate sequestration capacities. However, shrubland (24.28 t/h) and barren land (18.56 t/h), which are predominately located within the arid and semi-arid regions of the country, had the lowest carbon sequestration rates.

Future SOC stocks of the four projected SSPs ranged from 4.97 Pg to 5.3 Pg, indicating a substantial loss in SOC over the next few decades (Table 7.4). Nonetheless, it is worth noting that the rate of SOC change differed between classes, with the SOC stock of some classes increasing, while others declined. For instance, grassland, cropland, and shrubland displayed increases in SOC stocks across the SSPs (except for the very extreme SSP585). This high emission pathway showed a slight reduction for both grassland (-0.01 Pg) and shrubland (-0.03 Pg) SOC stock. The increases in SOC stocks recorded amongst these classes can be linked directly to their expansion in coverage, although their potential to accumulate more carbon may be hampered by their low to moderate sequestration rates compared to forest areas as shown in Table 7.5. Other classes witnessed significant losses in SOC stocks under the four pathways, with the greatest loss occurring within natural forested areas of the extreme SSP585 (-0.40 Pg) scenario. Once again this can be directly linked to the potential reduction in area coverage as shown from the projected land cover changes within South Africa (Table 7.3).

Table 7.4. Changes between current (2021) and future (2050) SOC stocks expressed in Petagrams (Pg), the values in parenthesis show the gain or loss in SOC stocks across different land cover types. The negative sign represents a loss.

Land cover type	**Current (Pg)**	**SSP126 (Pg)**	**SSP245 (Pg)**	**SSP370 (Pg)**	**SSP585 (Pg)**
Grassland	1.72	1.75 (0.03)	1.74 (0.02)	1.72 (0.00)	1.71 (-0.01)
Natural Forest	1.04	0.96 (-0.08)	0.87 (-0.17)	0.91 (-0.13)	0.64 (-0.40)
Commercial Forest	0.22	0.17 (-0.05)	0.14 (-0.08)	0.17 (-0.05)	0.12 (-0.10)
Cropland	0.87	0.89 (0.02)	1.01 (0.14)	0.94 (0.07)	1.03 (0.16)
Shrubland	1.12	1.13 (0.01)	1.12 (0.00)	1.12 (0.00)	1.09 (-0.03)
Barren land	0.35	0.21 (-0.14)	0.16 (-0.19)	0.12 (-0.23)	0.17 (-0.18)
Built-up vegetation	0.32	0.27 (-0.05)	0.17 (-0.15)	0.18 (-0.14)	0.21 (-0.11)
Total	**5.64**	**5.38**	**5.21**	**5.16**	**4.97**

Table 7.5. Mean SOC stocks density (t/h) for current and future predictions, the values in parenthesis shows the difference in mean density. The negative sign represents a loss.

Land cover type	Current (t/h)	SSP126 (t/h)	SSP245 (t/h)	SSP370 (t/h)	SSP585 (t/h)
Grassland	30.72	29.62 (-1.10)	27.01 (-3.71)	26.05 (-4.67)	26.42 (-4.30)
Natural Forest	45.01	51.22 (6.21)	51.14 (6.13)	48.65 (3.64)	49.25 (4.24)
Commercial Forest	46.95	52.50 (5.55)	53.05 (6.10)	50.55 (3.60)	50.64 (3.69)
Cropland	34.55	33.11 (-1.44)	30.45 (-4.10)	31.22 (-3.33)	29.53 (-5.02)
Shrubland	24.28	25.41 (1.13)	22.55 (-1.71)	21.65 (-2.63)	19.56 (-4.72)
Barren land	18.56	12.66 (-5.90)	13.25 (-5.31)	12.02 (-6.54)	10.11 (-8.45)
Built-up vegetation	38.32	39.44 (1.12)	38.89 (0.57)	36.44 (-1.88)	33.56-4.76)

7.4. Discussion

Climate, terrain, and land use changes have a substantial influence on SOC formulation, accumulation and dispersion (Zhao *et al.,* 2021; Yigini & Panagos 2016). In the wake of unforeseen climatic and land-use transformations, it has become necessary to anticipate and quantify the effects of future climate and land-use change scenarios on SOC stock inventories and adjust accordingly (Wiesmeier *et al.,* 2016). To provide an updated perspective on South Africa's SOC stock inventories and assist policymakers in coping with environmental and anthropogenic change, the study input simulated climatic, and land-use change scenarios into a deep learning framework to model spatial and temporal changes in SOC across South Africa's key land-uses.

7.4.1. Temporal assessment of South Africa's SOC stocks

The spatial variability of SOC remains fairly consistent between both current (2021) and future (2050) predictions, with greater SOC stocks prevalent in the north and south-east of the country as opposed to the arid regions in the west (Figure 7.4 and 7.5). The high SOC concentrations currently exhibited in these areas could be attributed to dense natural and exotic plantations which promote accelerated soil metabolism via continuous litterfall and dead matter, resulting in increased SOC accumulation (Odebiri *et al.,* 2022a; Kumar *et al.,* 2016).

Nonetheless, South African SOC stocks are projected to diminish, with total SOC stocks declining from 5.64 Pg in 2021, to between 4.97 Pg and 5.38 Pg by 2050, depending on emission rates (Table 7.4). This future decline in SOC stocks is consistent with other SOC projections (Caddeo *et al.,* 2019; Gray and Bishop, 2016; Wiesmeier *et al.,* 2016). For instance, both Caddeo *et al.,* (2019) and Wiesmeier *et al.,* (2016) predicted future SOC to decline within parts of Europe, with Wiesmeier *et al.,* (2016) documenting substantial SOC losses of between 11 and 16%. Meanwhile in the Southern Hemisphere, Gray and Bishop (2016) estimated that New South Wales in Australia will lose approximately 8.7 Mg ha^{-1} of SOC over the next 50 years because of climate change. This disruption in SOC stocks could be attributed to various climatic and land-use transformations over time, as the accumulation and storage of SOC is often dependent on an environment's exposure to specific topographic, climatic and land-use conditions (Lal, 2009). Coincidently, specific climatic extremes (such as elevated temperatures and lower rainfall) that are driven by global warming, may push ecosystems towards a lower SOC equilibrium (Padarian *et al.,* 2021; Melillo *et al.,* 2017; Crowther *et al.,* 2016). Specifically, we postulate that the dry semi-arid conditions, which are accompanied by extreme temperatures and prolonged periods of drought (a characteristic of the Northern, Western and Eastern Cape provinces) may over time expand towards the eastern interior of the country, significantly altering sensitive ecosystems and reducing SOC (Figures 7.4 and 7.5). For example, South Africa's eastern regions receive considerably more rainfall (> 600 mm) than the arid west (< 300 mm). With these disparities set to increase with global warming, lower rainfall will substantially affect soil moisture, vegetation density and litter decomposition in these areas, resulting in lower SOC sequestration (Odebiri *et al.,* 2020b). Moreover, relatively shallow sandy soils with little water retention capacity and wind erosion (dunes), may further constrain SOC accumulation within South Africa's western regions. This may be compounded by an increase in shrublands, which are more resilient to climate change but have lower SOC sequestration capacity (Figure 7.1).

Nevertheless, the biggest contributor to SOC loss is land-use disturbance (Garcia-Pausas *et al.,* 2007; Ramesh *et al.,* 2019; Were *et al.,* 2015). Although 45 % of South Africa's SOC has already been lost to intensive agriculture (Du Preez *et al.,* 2011), croplands are projected to expand by 5191.54 km^2 over the next 30 years (Table 7.3). An increase in large-scale intensive agriculture will significantly deplete SOC stocks (Lal, 2004; Olsson & Ardö, 2002). Earlier studies by Du Toit *et al.,* (1994), and Du Toit & Du Preez, (1995), highlighted the impact of agriculture on SOC stocks in South Africa, with warm dry regions, where soil temperatures would impact the microbial mineralization of carbon, becoming more susceptible to SOC loss. Consequently, as agricultural activities intensify and climate change results in regions becoming more arid, greater SOC loss is undoubtedly anticipated. Despite this, other studies (i.e. Yigini and Panagos , 2016; Lugato *et al.* 2014; ÁLvaro-Fuentes *et al.,* 2011) have projected SOC stocks to increase as a result of enhanced Net Primary Productivity (NPP). These differences suggest that projected SOC stocks might be location-specific and may be influenced by the predominant environmental drivers of SOC and management systems within that region (Jost *et al.*, 2021). Furthermore, the accuracy of these projections is dependent on the robustness of the data and modelling procedure used (Makridakis, 1993). Overall, the anticipated SOC stock losses could impact soil fertility and ecosystem resilience, further exacerbating climate change. Consequently, comprehensive land-use planning frameworks that are supported by soil monitoring programs with integrated early warning indicators are required to secure future SOC stocks and safeguard ecosystem longevity (Bagstad *et al.*, 2013).

7.4.2. Assessment of current and future SOC stocks across major South African land uses

An evaluation of current and future SOC stock distribution across South Africa major land uses was conducted. Results showed that SOC stock concentrations corresponded to the amount of land occupied by each class between the different temporal periods. For instance, grasslands and shrubland, which currently represent 54,33 % of South Africa's total land mass, accounted for more than 50% of the total SOC stock (5.64 Pg) in 2021. Meanwhile in 2050, despite overall SOC stocks decreasing to an average of 5.18 Pg across the different emission pathways, these regions are expected to expand their overall coverage to 55,68 % and account for approximately 55% of future SOC stocks (2.845 Pg) (Table 7.4). Grasslands are comprised of diverse grasses, graminoids and forbs that facilitate SOC production through continuous organic litter decomposition (Du Preez & Snyman, 1993; Mills & Fey, 2003). In addition, the deep-rooted underground storage of these ecosystems may shield future SOC stocks from the effects of climate change (Ward *et al.,* 2016). Conversely, shrublands, which predominately occupy South Africa's arid regions, are comprised of mostly Spekboom (*Portulacaria afra*) which enhanced its SOC sequestration capabilities (Mills *et al.*, 2015). These drought-resistant plants, with rapid growth rates and high litter outputs have the potential to sequester 168 t C ha^{-1} of SOC and cope with future climate variations (Mills and Fey, 2004). Similarly, croplands showed an increase in SOC stocks between 2021 and 2050 (Table 7.4). This increase in SOC could be tied to the increased use of fertilizer and irrigation to bolster productivity and facilitate food security (SANLC, 2020). In contrast, forested areas, vegetation in built-up areas and barren land showed a reduction in future SOC stocks (Table 7.4). This change could be the result of a reduction in overall coverage, as escalating urbanization and agricultural conversion may drive future land-use transformation (Figure 7.1, Table 7.3). For instance, according to an assessment by Global Forest Watch (https://www.globalforestwatch.org/) on natural forest and tree cover, South Africa lost approximately 11 200 hectares of natural forest between 2002 and 2020 to deforestation. Correspondingly, South Africa is set to lose an additional 14 810.09 km^2 of forest over the next few decades (Table 7.3), eliminating 0.265 Pg from the overall SOC pool.

To assist policymakers in adjusting SOC management frameworks, the study also investigated the temporal fluctuation of carbon sequestration rates between the different land cover types in South Africa. Most land cover classes, however, demonstrated a reduction in carbon sequestration potential across the different emission pathways over time (Table 7.5). For example, the SOC sequestration rate of grasslands will decrease from 30.72 t/h in 2021, to between 26.05t/h and 29.62t/h by 2050 (Table 7.5). This reduction in SOC sequestration rates within grasslands is likely the result of grassland degradation induced by increased livestock grazing. According to Dlamini *et al.,* (2014), SOC stocks in the highly grazed highlands of KwaZulu-Natal, South Africa, have declined by almost 90%. Unfortunately, a further 1.2 % to 4.2 % of grassland SOC supplies are expected to be lost to overgrazing globally (Dlamini *et al.*, 2014). Both natural and commercial forests, however, exhibited an increase in sequestration rates despite a reduction in overall coverage (Table 7.3 and 7.5). A longer rooting residency and the high canopy coverage of forests may have increased SOC sequestration capacity, which might have been facilitated through an accelerated soil metabolism from continuous litter fall and dead matter (Muchena, 2017). Additionally, climatic changes may have prompted an increase in net primary productivity within these ecosystems (Zhu *et al.*, 2007). However, SOC accumulation can only be sustained for a specific period due to a natural carbon storage threshold (Padarian *et al.*, 2021).

To safeguard SOC stocks, governments and policymakers need to adjust existing management protocols and policies to maintain and rehabilitate degraded ecosystems (Douglass *et al.*, 2011; Woomer, 1993). For instance, improved farming methods (such as crop rotation, agro-forestry, and fallow systems) as well as grassland regeneration through better management practices have been recommended as ways to increase sequestration rates (Ajani *et al.*, 2013; Bangroo *et al.*, 2013). Moreover, carbon accounting projects, such as the Kenyan Agricultural Carbon Project, may encourage sustainable and climate-friendly farming techniques that support soil carbon sequestration (Makambo & Kisaka, 2017).

7.4.3. Evaluation of Model performance

The DNN model used in this study performed well in mapping current (2021) SOC stocks at a national level, with an overall accuracy of $R^2 = 0.7102$ and an RMSE of 7.44 t/h. This is an improvement on the national SOC models previously produced by Venter *et al.*, (2021) ($R^2 = 0.659$) and Schultze & Schutte (2020) ($R^2 = 0.203$), which used the random forest algorithm and field-level SOC median calculation approaches. This difference is likely due to the DNN model's ability to learn and extract more representative characteristics from SOC data via its multiple hidden layers of neurons (Ma *et al.*, 2019). DNN can also successfully simulate the dynamic interrelationship between SOC and other variables, capturing any potential relationships (Yuan *et al.*, 2020). Lastly, the estimated total SOC stocks of 5.64 Pg produced by the DNN model is comparable to findings by Venter *et al.*, (2021). Venter *et al.*, (2021) calculated approximately 5.59 Pg of total SOC for South Africa. This difference of 0.05 Pg, may be caused by discrepancies in the total area mapped (about 1.18 to 1.2 million km^2), quality of soil data used, as well as the algorithms adopted.

The future SOC projections, however, displayed a reduction in accuracies (from $R^2 = 0.598$ – 0.643 and RMSE = 11.37 t/h – 13.56 t/h) across all four Shared Socio-economic Pathways (SSPs SSP126, SSP245 and SSP370) developed from the average of five different climatic models (i.e. CNRM-CM61-1, CanESM5, GFDL-ESM4, ACCESS-ESM1-5, and INM-CM5-0). The absence of spectral data (Sentinel 3 metrics) for the modelling of future SOC stocks, however, could be one of the most evident reasons for the decline in accuracy. The importance of spectral data to SOC mapping has been noted in the literature, particularly in the VIS-NIR wavelength range of the electromagnetic spectrum, which gives essential reflectance information on SOC and is regarded the most sensitive region to determine SOC content (Odebiri *et al.*, 2022a, Lin *et al.*, 2020, Bilgili *et al.*, 2010). This is also supported by our findings, which revealed that Sentinel-3 spectral metrics and their derived vegetation indices, such as Band 8 (665nm), NDVI, EVI, and Band 11 (708.25nm), were among the top 10 most important variables for current (2021) SOC stock distribution (Figure 7.3B). These spectral variables have been documented to be sensitive to vegetation properties, such as biomass and chlorophyll content (Odebiri *et al.*, 2021a; Zhang *et al.*, 2019). These variables are considered important in SOC mapping due to the direct correlation between vegetation and SOC concentration (Taghizadeh-Mehrjardi *et al.*, 2020). Nevertheless, the accuracies obtained for future SOC models may be considered acceptable, especially given that to the best of our knowledge, these projections provide the first ever glimpse into the future of South African SOC stocks, particularly at a national level.

7.5. Conclusion

To assist relevant stakeholders in coping with rapid environmental and anthropogenic change and to provide an updated perspective on South Africa's SOC stock inventories, we input simulated climatic, and land-use conversion scenarios into a deep neural network to model spatial and temporal changes in SOC across South Africa's key land-uses to the year 2050. Results demonstrated a reduction in overall SOC stocks over the next 28 years, with South Africa set to lose an average of 0.46 Pg of SOC to land-use and climate changes. In addition, South Africa is projected to lose an additional 14 810.09 km^2 of forest cover over the next few decades, significantly impacting SOC sequestration rates. Moreover, the SHAP technique identified Rainfall, NDVI, Band 8 (665 nm), Elevation, Land cover, Temperature, EVI, Slope, Band 11 (708.25 nm) and TWI to be among the top ten most important explanatory variables for SOC stock distribution. The knowledge generated by this study is critical for informing national policies, rehabilitation, restoration, and intervention efforts aimed at mitigating climate change and improving soil quality. Nevertheless, DL frameworks are still constrained by the need for big data and high computing power. Consequently, although the DNN model performed relatively well for a large-scale mapping endeavour, more detailed soil inventory data across different land cover categories are needed at a regional scale to supplement soil and pedological datasets and alleviate potential overestimation issues. Future research could benefit from investigating the capability of higher spatial resolution sensors, as well as the addition of new datasets for SOC projections.

7.6. Summary

The effects of simulated climate and land-cover change on SOC inventories across South Africa were modelled in this chapter until the year 2050. To model present SOC content, a digital soil mapping technique was employed in conjunction with a deep neural network (DNN). Following that, future SOC stocks were calculated using WorldClim global circulation models and a space-for-time substitution (SFTS) method for four common socio-economic emission paths. Depending on these emission rates, future SOC inventories are projected to decrease over the next few decades. Moreover, essential ecosystems which support overall SOC accumulation (such as forests) are projected to decrease in size by the year 2050, substantially impacting regional SOC sequestration capabilities. These findings offer insight into the existing and future state of South Africa's SOC stock inventories, as well as the impact of climate and land-use change. The final chapter contextualizes all of the research findings while proposing recommendations for future research, having acknowledged the importance of the conclusions and deductions offered within this thesis.

Chapter Eight:
The application of deep learning for remote sensing of soil organic carbon stocks distribution in South Africa: A Synthesis

8.1. Introduction

Soil organic carbon (SOC) is the world's largest terrestrial carbon store, and its response to land use and management makes it an attractive prospect for carbon sequestration (Padarian *et al.,* 2021; Wang *et al.,* 2018; Minasny *et al.,* 2017). Soil organic carbon as a proxy for soil organic matter content is also a major determinant of soil quality and soil fertility (Zhao *et al.,* 2021; Batjes *et al.,* 2020). Its sequestration can contribute substantially to climate change mitigation as current estimates suggests that soils to 1 m depth hold approximately 74% of the total terrestrial carbon stocks (Köchy e*t al.,* 2015; Batjes *et al.,* 2017). Consequently, many global initiatives to mitigate climate change and land degradation are increasingly relying on accurate and detailed SOC inventories at (sub) national levels (Sahoo *et al*., 2019; Mishra *et al*., 2019). In addition, SOC stocks is recognized as one of the three Land Degradation Neutrality (LDN) indicators used by the United Nations Convention to Combat Desertification (UNCCD), necessitating an agreement among UNCCD stakeholders to report on SOC stock trends at regular time intervals (UNCCD, 2019). Given these significances, and the realization that even marginal SOC stocks increase translates to globally relevant magnitude of carbon, quantifying the spatial heterogeneity of SOC stocks at national and global scales, as well as its environmental controllers is necessary (Mishra *et al.,* 2019). Besides, sufficient understanding of national and global SOC estimates, trends, and distribution will be crucial for devising appropriate SOC management methods to improve carbon assimilation and meet the IPCC and Kyoto Protocol targets (IPCC 2016; Ndalowa 2014).

Based on the recent advances in digital soil mapping (DSM), which involve the use of remote sensing and other ancillary data together with spatial explicit models, modelling and monitoring of SOC at regular intervals is now achievable and less tedious when compared to traditional methods of SOC determination (Heuvelink *et al.,* 2021; Owusu *et al.,* 2020; Padarian *et al.,* 2019). The multiplication of different image datasets characterized by low-moderate-high temporal, spatial and spectral resolutions, presents an opportunity to map SOC at different spatial extents (Hamida *et al.,* 2018; Odindi *et al.,* 2016; Mutanga *et al.,* 2015). In addition, the use of advanced modelling algorithms like deep learning (DL) with proven ability to extract invariant and abstract features from image datasets leading to better discrimination capabilities, could permit a more reliable quantification of different environmental properties including SOC (Wadoux *et al.*, 2019; Zhang *et al.*, 2019; Litjens *et al.*, 2017). DL models can accurately approximate the complicated non-linear relationship between SOC and environmental covariates, thus capturing the potential association between them (Yuan *et al.*, 2020). Therefore, this study explored deep learning-based strategies for a national scale analysis of remote sensing data to predict SOC stocks distribution across South Africa, which then creates an effective framework for continuous national and global monitoring and management of soil organic carbon.

8.2. Conclusions

The essence of this study was to explore the utility and performance of deep learning (DL) approaches in concert with new generation remotely sensed data and environmental variables in estimating and mapping South Africa's soil organic carbon (SOC) stocks. The results of this study have shown that the use of remotely sensed environmental variables, new generation multispectral sensors and image processing techniques, integrated with advance non-parametric algorithms like DL, can accurately improve SOC estimations, particularly at a large spatial scale. Based on these findings, the following conclusions are drawn:

a) Based on the literature, remote sensing (RS) is a great tool for large-scale SOC mapping, a previously difficult task using typical laboratory SOC determination methods. The value and performance of RS data are also highly dependent on the algorithm used. Deep learning (DL) strategies based on neural networks are superior to geostatistical and other traditional machine learning (ML) models in terms of implementation and utility. Although DL models have had some success with hyperspectral RS data, their usage for multispectral data analytics is still in its infancy and requires further refinement. Furthermore, its application for SOC mapping, a vital climate change mitigation measure, has seen little progress, particularly in Africa.

b) The free and widely available Sentinel-3 Ocean and Land Colour Instrument (OLCI) multispectral sensor, with higher spectral resolution (21 bands between 400-1020nm), shorter revisit time (less than 2 days) large swath width (1270 km) and improved signal-to-noise ratios, provides an invaluable primary data source for accurate country-scale SOC estimation, especially in data-scarce environments. Moreover, the DNN model outperformed other traditional machine learning (ML) models, with NDVI, Red band

8 (665 nm), Red-edge band 11 (708.25 nm), EVI, and RVI being the best explanatory factors for SOC stock distribution in South Africa.

c) The distribution of SOC stocks across seven key South African land uses revealed that SOC accumulation is directly related to landscape size, with grassland landscapes having the highest SOC deposition and urban vegetation having the lowest. However, the mean SOC density of forest regions, both natural and commercial, revealed the highest sequestration rates.

d) The hybrid DL model (Concrete Autoencoder-Deep Neural Network (CAE-DNN)) which combines the integration of spectral bands and vegetation indices with multiple environmental variables (i.e. climate, topography and land use) can be used to improve the prediction of SOC stocks across the major nine South African biomes. The CAE-DNN optimal variable selection technique and regression provides a better SOC retrieval accuracy improvement compared to popular classical ML feature selection strategy. Results from analysis also indicated that the accumulation of SOC in different biome is directly proportionate to the size of the biome, with the exception of the grassland biome whose SOC accumulation is greater than the Savanna biome, despite the latter being larger in size. The forest and Indian ocean coastal belt biomes, despite their smaller footprint showed the highest sequestration rates, while the desert biome showed the least.

e) Future SOC projections and potential change detections are viable through the combination of DL-based Digital soil mapping (DSM) strategy together with the principle of space-for-time substitution (SFTS), where current covariates are exchanged with projected future covariates (climate and land use) in the model to forecast SOC stocks. The general decline in the projected SOC stocks from the current accumulation together with the complexities associated with different land use classes evaluated as obtained from the result, is a testament to the inadequacy and efficiency of the current SOC management strategies and policies in South Africa.

f) Overall, all the categories of environmental variables used in this study including spectral information, topo-climate and land use metrics were all important to SOC stocks simulation. The most important 20 covariates for mapping and modelling SOC stocks in South Africa were Rainfall, NDVI, Sentinel-3 Red band 8 (665 nm), Sentinel-3 Red-edge band 11 (708.25 nm), Elevation, Land use, Temperature, TWI, EVI, RVI, Slope, SAVI, DVI, Sentinel-3 NIR band 13 (761.25 nm), Sentinel-3 NIR band 17 (865 nm), Sentinel-3 NIR band 18 (885 nm), Sentinel-3 NIR band 19 (900 nm), Sentinel-3 Blue band 7 (620 nm), Sentinel-3 Blue band 6 (560 nm) and General Curvature. This study underscores the utility of DL-based remotely sensed data in providing invaluable data set for regional and global SOC stocks accounting.

8.3. Challenges and limitations

Although deep learning (DL) models have shown success in a variety of remote sensing applications, including SOC modelling, they face a number of challenges that limit their utility. Following a thorough review, we established that large sample size requirements, computational time, interpretability, end-user technical know-how, large storage capacity requirements, and the tendency to over-fit, are some of the factors limiting the use of DL architectures. Moreover, there is variation in the results/accuracy where these strategies have been utilized. This could be attributed to discrepancies in calibration data selection as well as SOC variation in the study areas, as SOC is a dynamic phenomenon that varies by region. As such, there is currently no widely acknowledged calibration procedure for SOC retrieval (Lamichhane *et al.,* 2019).

The quality of soil profile data used in this study was a major challenge due to the use of legacy soil inventory data which span across different times and seasons with different sampling sources, depth, and the method of SOC concentration determination. In addition, some predictor variables may not accurately represent current conditions. For example, bio-climatic factors (i.e. rainfall and temperature) used in this study represent the average values from the year 1970 to 2000. Also, these variables are interpolated datasets from the global weather stations and developing countries particularly in Africa, do not have reliable weather stations, thereby producing a generalized description of environmental variability. Similarly, spatial information from various sources (in our case, Sentinel-3 — 300 m, SRTM DEM — 90 m, WorldClim — 1km and Land use map — 20 m) do not have the same resolution. Resampling datasets into the same spatial resolution may lead to uncertainties. All these limitations and challenges will may reduce the quality of the simulation process by reducing overall accuracy. Notwithstanding the spatial uncertainty of the target SOC data as aforementioned, the DNN model used in this study provided a decent outcome, demonstrating its robustness in modelling complicated data, particularly at a national scale.

8.4. Recommendation and future prospects

To the best of our knowledge, this is the first deep learning-based remote sensing modelling of soil organic carbon (SOC) stocks in South Africa. The study produced the best national SOC distribution maps with the highest accuracy when compared to other studies. It is also the first study to include potential future SOC maps (2050), revealing a general decline in SOC stocks from current levels which is critical for evaluating current SOC management practices and policies. The study findings also highlighted the role of SOC in the global carbon cycle, which could serve as a foundation for conservation strategies aimed at facilitating the effective and long-term use of soil in various settings. In addition, the findings of this study demonstrated that deep learning-based remote sensing methodologies provide a reliable and robust primary data source, as well as a powerful analytical algorithm for quantifying and mapping SOC stocks. Similarly, the results also provide critical information to the remote sensing community, ecologists, and environmentalists about the use of free and easily available sensors for effective SOC stock monitoring. This is particularly significant in data-poor regions, where the adoption of high-resolution aerial and hyperspectral sensors is still difficult due to the related costs. This study, therefore, suggests the following recommendations and future prospect:

a) Because of its complexity, DL models necessitate a high sample size. However, most studies are constrained by small field (in situ) samples. Furthermore, cloud cover and insufficient ground data limit the amount of samples in remote sensing datasets, resulting in missing satellite data. The Transfer Learning technique proposed by Goodfellow *et al.,* (2016) could be used to overcome this problem. This approach works by adjusting DL model parameters of a formally trained large dataset with smaller samples for optimum implementation on the new task. Additionally, while we recognize that DL models require a significant amount of computational power and large datasets to function properly, recent advances in computing power and storage, as well as innovations like cloud computing and analytical APIs like Google Earth Engine, may overcome these constraints.

b) Although this study solely made use of Deep Neural Network (DNN) model for regression and Concreate Autoencoder (CAE) for feature selection, future studies can compare the trade-off between the DNN and other powerful DL architecture such as Convolutional Neural Network (CNN) and Deep Belief Networks (DBN) in terms of accuracy and time of training.

c) Although the Sentinel-3 data utilized in this study yielded relatively satisfactory results, future studies should compare and evaluate Sentinel 3's performance against other freely available multispectral sensors with higher spatial resolution, such as Sentinel-2, Landsat-8, and Landsat-9. As a result, end-users may be able to select the appropriate image datasets for different mapping scales. However, users should note that while employing DL, the higher the image spatial resolution and area coverage, the more computing power (such as GPUs) is required.

d) Future research could include the use of Light Detection and Ranging (LIDAR) and the freely available Sentinel-1 Synthetic Aperture Radar (SAR), both which have yet to be fully explored for DL-based SOC retrieval. LIDAR provides higher-spatial resolution data than SRTM DEMs, which could be useful for DL models. Nonetheless, the high LIDAR cost, the need for ground data for calibration, and impracticability in some remote areas, particularly in Africa, are some of the challenges to its widespread use. Sentinel-1, on the other hand, can be used as a substitute for LIDAR in DL RS-based SOC mapping due to its low cost and relatively high resolution. Similarly, the application of DL-based unmanned aerial system (UAS) platforms for SOC mapping has yet to be investigated. The UAS platforms are cheaper and allow for excellent data collection settings. However, because it is difficult to employ UAS over a vast spatial extent, its usage may be limited by coverage.

e) Subsequent studies can incorporate the fusion of different remotely sensed data types to improve accuracy. In remote sensing applications, image fusion is typically used to create a single image with higher spectral and spatial resolution. For example, combining Sentinel-2 and Sentinel-3 to improve accuracy by taking full advantage of their spatial and spectral resolutions respectively. Similarly, SAR imaging fused with any of the freely accessible multispectral sensors could improve SOC stock prediction results. SAR imagery is made up of two satellites that carry C-band synthetic aperture radar sensors and are very sensitive to several soil qualities, including SOC. SAR also offer continuous images without being influenced by clouds, a typical issue with multispectral sensors.

f) This study's future SOC estimates (2050) were made using projected climate and land use, as well as topographical metrics, without using spectral information/vegetation indices like NDVI, which has been shown to be crucial for SOC mapping. Subsequent studies can consider including relevant spectral derived vegetation indices like NDVI by leverage on their present and historical conditions to project to the desired future and thereafter used as part of the covariates to perform future potential SOC modelling, hence improving accuracy.

g) Although the DNN model performed quite well in this study, despite the spatial uncertainty of the legacy soil data used, we recommend that the South Africa government and research agencies fund standardized soil inventory schemes where quality data unique to distinct land use types and biomes is gathered and made available for SOC modelling efforts.

h) Finally, while we applaud and encourage efforts to rehabilitate degraded ecological areas, we caution against initiatives to convert other naturally existing land uses into forests in order to boost SOC sequestration in response to the global carbon market. Altering the natural status of other important ecosystems could be harmful to biodiversity and counterproductive. For instance, converting natural grassland regions to forest areas may not be ideal, because grasslands, unlike other landscapes, store most of their accumulated carbon in the soil, making them more tolerant to disasters like wildfires.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., & Acharya, U. R. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*.

Abid, A., Balin, M. F., & Zou, J. (2019). Concrete autoencoders for differentiable feature selection and reconstruction. *arXiv preprint arXiv:1901.09346*.

Afshar, F. A., Ayoubi, S., & Jalalian, A. (2010). Soil redistribution rate and its relationship with soil organic carbon and total nitrogen using 137Cs technique in a cultivated complex hillslope in western Iran. *Journal of Environmental Radioactivity*, *101*(8), 606-614.

Ajani, E., Mgbenka, R., & Okeke, M. (2013). Use of indigenous knowledge as a strategy for climate change adaptation among farmers in sub-Saharan Africa: implications for policy.

ÁLvaro-Fuentes, J., Easter, M., Cantero-Martinez, C., & Paustian, K. (2011). Modelling soil organic carbon stocks and their changes in the northeast of Spain. *European journal of soil science*, *62*(5), 685-695.

Amanuel, W., Yimer, F., & Karltun, E. (2018). Soil organic carbon variation in relation to land use changes: the case of Birr watershed, upper Blue Nile River Basin, Ethiopia. *Journal of Ecology and Environment*, *42*(1), 1-11.

Ampleman, M. D., Crawford, K. M., & Fike, D. A. (2014). Differential soil organic carbon storage at forb-and grass-dominated plant communities, 33 years after tallgrass prairie restoration. *Plant and soil*, *374*(1), 899-913.

Angelopoulou, T., Balafoutis, A., Zalidis, G., & Bochtis, D. (2020). From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation—A Review. *Sustainability*, *12*(2), 443.

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., & Bochtis, D. (2019). Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing*, *11*(6), 676.

Arletti, F., Persiano, S., Bertola, M., Parajka, J., Blöschl, G., & Castellarin, A. (2021). Recent spatio-temporal dynamics of floods of record across Europe. EGU General Assembly Conference Abstracts,

Arogoundade, A. M., Odindi, J., & Mutanga, O. (2019). Modelling Parthenium hysterophorus invasion in KwaZulu-Natal province using remotely sensed data and environmental variables. *Geocarto International*, 1-16.

Aryal, D. R., De Jong, B. H. J., Mendoza-Vega, J., Ochoa-Gaona, S., & Esparza-Olguín, L. (2017). Soil organic carbon stocks and soil respiration in tropical secondary forests in Southern Mexico. In *Global soil security* (pp. 153-165). Springer.

Atela, J. O. (2012). The politics of Agricultural carbon finance: The case of the Kenya Agricultural Carbon Project.

Ayoubi, S., Shahri, A. P., Karchegani, P. M., & Sahrawat, K. L. (2011). Application of artificial neural network (ANN) to predict soil organic matter using remote sensing data in two ecosystems. *Biomass and remote sensing of biomass*, 181-196.

Bagstad, K. J., Semmens, D. J., Waage, S., & Winthrop, R. (2013). A comparative assessment of decision-support tools for ecosystem services quantification and valuation. *Ecosystem Services*, *5*, 27-39.

Baldock, J., Wheeler, I., McKenzie, N., & McBrateny, A. (2012). Soils and climate change: potential impacts on carbon stocks and greenhouse gas emissions, and future research for Australian agriculture. *Crop and Pasture Science*, *63*(3), 269-283.

Balesdent, J., Basile-Doelsch, I., Chadoeuf, J., Cornu, S., Derrien, D., Fekiacova, Z., & Hatté, C. (2018). Atmosphere–soil carbon transfer as a function of soil depth. *Nature*, *559*(7715), 599-602.

Bangroo, S., Ali, T., Mahdi, S. S., Najar, G., & Sofi, J. (2013). Carbon and greenhouse gas mitigation through soil carbon sequestration potential of adaptive agriculture and agroforestry systems. *Range Management and Agroforestry*, *34*(1), 1-11.

Baret, F., & Guyot, G. (1991). Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote Sensing of Environment*, *35*(2-3), 161-173.

Batjes, N. H. (1996). Total carbon and nitrogen in the soils of the world. *European journal of soil science*, *47*(2), 151-163.

Batjes, N. H., Ribeiro, E., & Oostrum, A. v. (2020). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, *12*(1), 299-320.

Batjes, N. H., Ribeiro, E., Van Oostrum, A., Leenaars, J., Hengl, T., & Mendes de Jesus, J. (2017). WoSIS: providing standardised soil profile data for the world. *Earth System Science Data*, *9*(1), 1-14.

Beecham, S., Razzaghmanesh, M., Bustami, R., & Ward, J. (2019). The role of green roofs and living walls as WSUD approaches in a dry climate. In *Approaches to Water Sensitive Urban Design* (pp. 409-430). Elsevier.

Bhandari, S., & Bam, S. (2013). Comparatives study of soil organic carbon (soc) under forest, cultivated and barren land: A case of Chovar village, Kathmandu. *NJST*, *14*(2), 103-108.

Bhunia, G. S., Kumar Shit, P., & Pourghasemi, H. R. (2017). Soil organic carbon mapping using remote sensing techniques and multivariate regression model. *Geocarto Int.*, *34*, 1-12.

Bilgili, A. V., Van Es, H., Akbas, F., Durak, A., & Hively, W. (2010). Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. *Journal of Arid Environments*, *74*(2), 229-238.

BODAGHABADI, M. B., MARTÍNEZ-CASASNOVAS, J., Salehi, M. H., Mohammadi, J., BORUJENI, I. E., Toomanian, N., & Gandomkar, A. (2015). Digital soil mapping using artificial neural networks and terrain-related attributes. *Pedosphere*, *25*(4), 580-591.

Böhner, J., Koethe, R., Conrad, O., Gross, J., Ringeler, A., & Selige, T. (2001). Soil regionalisation by means of terrain analysis and process parameterisation. *Soil classification*(7), 213.

Böhner, J., & Selige, T. (2006). Spatial prediction of soil attributes using terrain analysis and climate regionalisation.

Bond, W. J., Stevens, N., Midgley, G. F., & Lehmann, C. E. (2019). The trouble with trees: afforestation plans for Africa. *Trends in ecology & evolution*, *34*(11), 963-965.

Boon, R., Cockburn, J., Govender, N., Ground, L., Slotow, R., Mclean, C., Douwes, E., Rouget, M., & Roberts, D. (2016). Managing a threatened savanna ecosystem (KwaZulu-Natal Sandstone Sourveld) in an urban biodiversity hotspot: Durban, South Africa. *Bothalia-African Biodiversity & Conservation*, *46*(2), 1-12.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Broderick, D. E., Frey, K. E., Rogan, J., Alexander, H. D., & Zimov, N. S. (2015). Estimating upper soil horizon carbon stocks in a permafrost watershed of Northeast Siberia by integrating field measurements with Landsat-5 TM and WorldView-2 satellite data. *GIScience & Remote Sensing*, *52*(2), 131-157.

Broomhead, D. S., & Lowe, D. (1988). *Radial basis functions, multi-variable functional interpolation and adaptive networks*.

Caddeo, A., Marras, S., Sallustio, L., Spano, D., & Sirca, C. (2019). Soil organic carbon in Italian forests and agroecosystems: Estimating current stock and future changes with a spatial modelling approach. *Agricultural and Forest Meteorology*, *278*, 107654.

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.

Chen, C.-H., Kung, H.-Y., & Hwang, F.-J. (2019). Deep learning techniques for agronomy applications. In: Multidisciplinary Digital Publishing Institute.

Chen, H., Liu, Z., Gu, J., Ai, W., Wen, J., & Cai, K. (2018). Quantitative analysis of soil nutrition based on FT-NIR spectroscopy integrated with BP neural deep learning. *Analytical Methods*, *10*(41), 5004-5013.

Chen, S., Xu, D., Li, S., Ji, W., Yang, M., Zhou, Y., Hu, B., Xu, H., & Shi, Z. (2020). Monitoring soil organic carbon in alpine soils using in situ vis-NIR spectroscopy and a multilayer perceptron. *Land Degradation & Development*, *31*(8), 1026-1038.

Chen, X., Zhang, D., Liang, G., Qiu, Q., Liu, J., Zhou, G., Liu, S., Chu, G., & Yan, J. (2015). Effects of precipitation on soil organic carbon fractions in three subtropical forests in southern China. *J plant ecol*, *9*(1), 10-19.

Chen, Y., Lin, Z., Zhao, X., Wang, G., & Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(6), 2094-2107.

Chi, Y., Shi, H., Zheng, W., & Sun, J. (2018). Simulating spatial distribution of coastal soil carbon content using a comprehensive land surface factor system based on remote sensing. *Science of The Total Environment*, *628*, 384-399.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Ciais, P., Bombelli, A., Williams, M., Piao, S., Chave, J., Ryan, C., Henry, M., Brender, P., & Valentini, R. (2011). The carbon balance of Africa: synthesis of recent research studies. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *369*(1943), 2038-2057.

CireşAn, D., Meier, U., Masci, J., & Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural networks*, *32*, 333-338.

Cockburn, J., Rouget, M., Slotow, R., Roberts, D., Boon, R., Douwes, E., O'Donoghue, S., Downs, C. T., Mukherjee, S., & Musakwa, W. (2016). How to build science-action partnerships for local land-use planning and management: lessons from Durban, South Africa. *Ecology and Society*, *21*(1).

Conant, R. T., Paustian, K., & Elliott, E. T. (2001). Grassland management and conversion into grassland: effects on soil carbon. *Ecological Applications*, *11*(2), 343-355.

Conant, R. T., Ryan, M. G., Ågren, G. I., Birge, H. E., Davidson, E. A., Eliasson, P. E., Evans, S. E., Frey, S. D., Giardina, C. P., & Hopkins, F. M. (2011). Temperature and soil organic matter decomposition rates–synthesis of current knowledge and a way forward. *Global Change Biology, 17*(11), 3392-3404.

Cong, R.-G., & Brady, M. (2012). The interdependence between rainfall and temperature: copula analyses. *Sci. World J., 2012*.

Cox, B. (2001). The biogeographic regions reconsidered. *Journal of Biogeography, 28*(4), 511-523.

Crowther, T. W., Todd-Brown, K. E., Rowe, C. W., Wieder, W. R., Carey, J. C., Machmuller, M. B., Snoek, B., Fang, S., Zhou, G., & Allison, S. D. (2016). Quantifying global soil carbon losses in response to warming. *Nature*, *540*(7631), 104-108.

CSIR, N. (2006). Climatic Future for Durban. *CSIR, Durban*.

Dai, F., Zhou, Q., Lv, Z., Wang, X., & Liu, G. (2014). Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. *Ecological Indicators*, *45*, 184-194.

Daniel, K., Tripathi, N., & Honda, K. (2003). Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). *Soil Research*, *41*(1), 47-59.

Das, S., & Sarkar, R. (2019). Predicting the land use and land cover change using Markov model: A catchment level analysis of the Bhagirathi-Hugli River. *Spatial Information Research*, *27*(4), 439-452.

Davy, M., & Koen, T. (2014). Variations in soil organic carbon for two soil types and six land uses in the Murray Catchment, New South Wales, Australia. *Soil Research*, *51*(8), 631-644.

de Araujo Barbosa, C. C., Atkinson, P. M., & Dearing, J. A. (2015). Remote sensing of ecosystem services: a systematic review. *Ecological Indicators*, *52*, 430-443.

De Deyn, G. B., Cornelissen, J. H., & Bardgett, R. D. (2008). Plant functional traits and soil carbon sequestration in contrasting biomes. *Ecology letters*, *11*(5), 516-531.

Deering, D. (1975). Measuring" forage production" of grazing units from Landsat MSS data. Proceedings of the Tenth International Symposium of Remote Sensing of the Envrionment,

Deng, C., Huang, G., Xu, J., & Tang, J. (2015). Extreme learning machines: new trends and applications. *Science China information sciences*, *58*(2), 1-16.

Department of Environmental Affairs 2017.National Terrestrial Carbon Sinks Assessment. *Department of Environmental Affairs, Pretoria, South Africa*

Desertification), U. U. N. C. t. C. ((2019)). *Report of the Conference of the Parties on its fourteenth session, held in New Delhi, India, from 2 to September 13, 2019. https://www.unccd.int/sites/default/files/sessions/ documents/2019-12/ICCD_COP%2814%29_23_Add.1-1918355E. pdf (accessed on January 2, 2020).*

Di Noia, A., & Hasekamp, O. P. (2018). Neural networks and support vector machines and their application to aerosol and cloud remote sensing: a review. In *Springer Series in Light Scattering* (pp. 279-329). Springer.

Dickson, B., & Kapos, V. (2012). Biodiversity monitoring for REDD+. *Current Opinion in Environmental Sustainability*, *4*(6), 717-725.

Dietz, W. H. (2020). Climate change and malnutrition: we need to act now. *The Journal of clinical investigation*, *130*(2), 556-558.

Ding, S., Zhao, H., Zhang, Y., Xu, X., & Nie, R. (2015). Extreme learning machine: algorithm, theory and applications. *Artificial Intelligence Review*, *44*(1), 103-115.

Diversity, S. o. t. C. o. B. (2009). Review of the Literature on the Links Between Biodiversity and Climate Change: Impacts, adaptation, and mitigation.

Dlamini, P., Chivenge, P., Manson, A., & Chaplot, V. (2014). Land degradation impact on soil organic carbon and nitrogen stocks of sub-tropical humid grasslands in South Africa. *Geoderma, 235*, 372-381.

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.

Dong, Z., Zhou, Z., Li, Z., Liu, C., Huang, P., Liu, L., Liu, X., & Kang, J. (2018). Convolutional neural networks based on RRAM devices for image recognition and online learning tasks. *IEEE Transactions on Electron Devices, 66*(1), 793-801.

Dotto, A. C., Dalmolin, R. S. D., ten Caten, A., & Grunwald, S. (2018). A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. *Geoderma, 314*, 262-274.

Douglass, L. L., Possingham, H. P., Carwardine, J., Klein, C. J., Roxburgh, S. H., Russell-Smith, J., & Wilson, K. A. (2011). The effect of carbon credits on savanna land management and priorities for biodiversity conservation. *PLoS One, 6*(9), e23843.

Dovey. (2014). Current carbon stock estimation capability for South African commercial forest plantations. Report produced for the Department of Environmental Affiars.

Du Preez, C., & Snyman, H. (1993). Research note: Organic matter content of a soil in a semi-arid climate with three long-standing veld conditions.

Du Preez, C. C., Van Huyssteen, C. W., & Mnkeni, P. N. (2011). Land use and soil organic matter in South Africa 2: A review on the influence of arable crop production. *South African Journal of Science, 107*(5), 1-8.

Du Toit, M., & Du Preez, C. (1995). Effect of cultivation on the nitrogen fertility of selected dryland soils in South Africa. *South African Journal of Plant and Soil, 12*(2), 73-81.

Du Toit, M., Du Preez, C., Hensley, M., & Bennie, A. (1994). Effek van bewerking op die organiese materiaalinhoud van geselekteerde droëlandgronde in Suid-Afrika. *South African Journal of Plant and Soil, 11*(2), 71-79.

Egoh, B. N., Reyers, B., Rouget, M., & Richardson, D. M. (2011). Identifying priority areas for ecosystem service management in South African grasslands. *Journal of environmental management, 92*(6), 1642-1650.

Ekblad, A., & Bastviken, D. (2019). Deforestation releases old carbon. *Nature Geoscience, 12*(7), 499-500.

El-Tantawi, A. M., Bao, A., Chang, C., & Liu, Y. (2019). Monitoring and predicting land use/cover changes in the Aksu-Tarim River Basin, Xinjiang-China (1990–2030). *Environmental monitoring and assessment, 191*(8), 1-18.

Ellis, E. C., Klein Goldewijk, K., Siebert, S., Lightman, D., & Ramankutty, N. (2010). Anthropogenic transformation of the biomes, 1700 to 2000. *Global Ecology and Biogeography, 19*(5), 589-606.

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020). Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing, 12*(14), 2234.

Esri, A. D. (2011). Release 10. *Documentation Manual. Redlands, CA, Environmental Systems Research Institute*.

Ethekwini municipality and Wildland's conservation trust. (2014). Community, Climate and Biodiversity Standard Project Design Document.

Falahatkar, S., Hosseini, S. M., Ayoubi, S., & Salmanmahiny, A. (2016). Predicting soil organic carbon density using auxiliary environmental variables in northern Iran. *Archives of Agronomy and Soil Science*, *62*(3), 375-393.

Fidencio, P. H., Poppi, R. J., & de Andrade, J. C. (2002). Determination of organic matter in soils using radial basis function networks and near infrared spectroscopy. *Analytica Chimica Acta*, *453*(1), 125-134.

Fiener, P., Dlugoß, V., & Van Oost, K. (2015). Erosion-induced carbon redistribution, burial and mineralisation—Is the episodic nature of erosion processes important? *Catena*, *133*, 282-292.

Fissore, C., Dalzell, B., Berhe, A., Voegtle, M., Evans, M., & Wu, A. (2017). Influence of topography on soil organic carbon dynamics in a Southern California grassland. *Catena*, *149*, 140-149.

Florinsky, I. (2016). *Digital terrain analysis in soil science and geology*. Academic Press.

Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PLoS One*, *12*(1), e0170478.

Gao, Q., Zribi, M., Escorihuela, M. J., & Baghdadi, N. (2017). Synergetic use of Sentinel-1 and Sentinel-2 data for soil moisture mapping at 100 m resolution. *Sensors*, *17*(9), 1966.

Gara, T. W., Murwira, A., & Ndaimani, H. (2016). Predicting forest carbon stocks from high resolution satellite data in dry forests of Zimbabwe: exploring the effect of the red-edge band in forest carbon stocks estimation. *Geocarto International*, *31*(2), 176-192.

Garcia-Pausas, J., Casals, P., Camarero, L., Huguet, C., Sebastia, M.-T., Thompson, R., & Romanya, J. (2007). Soil organic carbon storage in mountain grasslands of the Pyrenees: effects of climate and topography. *Biogeochemistry*, *82*(3), 279-289.

Gautam, R., Panigrahi, S., Franzen, D., & Sims, A. (2011). Residual soil nitrate prediction from imagery and non-imagery information using neural network technique. *Biosystems engineering*, *110*(1), 20-28.

Georgiou, K., Malhotra, A., Wieder, W. R., Ennis, J. H., Hartman, M. D., Sulman, B. N., Berhe, A. A., Grandy, A. S., Kyker-Snowman, E., & Lajtha, K. (2021). Divergent controls of soil organic carbon between observations and process-based models. *Biogeochemistry*, *156*(1), 5-17.

Ghimire, P., Bhatta, B., Pokhrel, B., Kafle, G., & Paudel, P. (2018). Soil organic carbon stocks under different land uses in Chure region of Makawanpur district, Nepal. *SAARC Journal of Agriculture*, *16*(2), 13-23.

Gholizadeh, A., Žižala, D., Saberioon, M., & Borůvka, L. (2018). Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sensing of Environment*, *218*, 89-103.

Gitelson, A. A., & Merzlyak, M. N. (1998). Remote sensing of chlorophyll concentration in higher plant leaves. *Advances in Space Research*, *22*(5), 689-692.

Goldstein, A., Turner, W. R., Spawn, S. A., Anderson-Teixeira, K. J., Cook-Patton, S., Fargione, J., Gibbs, H. K., Griscom, B., Hewson, J. H., & Howard, J. F. (2020). Protecting irrecoverable carbon in Earth's ecosystems. *Nature Climate Change*, *10*(4), 287-295.

Gomes, A. L., Revermann, R., Gonçalves, F. M., Lages, F., Aidar, M. P., Mostajo, G. A. S., & Finckh, M. (2021). Suffrutex grasslands in south-central Angola: belowground biomass, root structure, soil characteristics and vegetation dynamics of the 'underground forests of Africa'. *Journal of Tropical Ecology*, *37*(3), 136-146.

González-Roglich, M., Swenson, J. J., Villarreal, D., Jobbágy, E. G., & Jackson, R. B. (2015). Woody plant-cover dynamics in Argentine savannas from the 1880s to 2000s: the interplay of encroachment and agriculture conversion at varying scales. *Ecosystems*, *18*(3), 481-492.

Gonzalez, P., Neilson, R. P., Lenihan, J. M., & Drapek, R. J. (2010). Global patterns in the vulnerability of ecosystems to vegetation shifts due to climate change. *Global Ecology and Biogeography*, *19*(6), 755-768.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems,

Gray, J. M., & Bishop, T. F. (2016). Change in soil organic carbon stocks under 12 climate change projections over New South Wales, Australia. *Soil Science Society of America Journal*, *80*(5), 1296-1307.

Grebner, D. L., Bettinger, P., & Siry, J. P. (2012). *Introduction to forestry and natural resources*. Academic Press.

Griscom, B. W., Adams, J., Ellis, P. W., Houghton, R. A., Lomax, G., Miteva, D. A., Schlesinger, W. H., Shoch, D., Siikamäki, J. V., & Smith, P. (2017). Natural climate solutions. *Proceedings of the National Academy of Sciences*, *114*(44), 11645-11650.

Gruszczyński, S. (2019). Prediction of soil properties with machine learning models based on the spectral response of soil samples in the near infrared range. *Soil Science Annual*, *70*(4), 298-313.

Guan, S., An, N., Zong, N., He, Y., Shi, P., Zhang, J., & He, N. (2018). Climate warming impacts on soil organic carbon fractions and aggregate stability in a Tibetan alpine meadow. *Soil Biology and Biochemistry*, *116*, 224-236.

Guo, L., Fu, P., Shi, T., Chen, Y., Zhang, H., Meng, R., & Wang, S. (2020). Mapping field-scale soil organic carbon with unmanned aircraft system-acquired time series multispectral images. *Soil and Tillage Research*, *196*, 104477.

Gupta, D., Prasad, R., Srivastava, P., & Islam, T. (2016). Nonparametric Model for the Retrieval of Soil Moisture by Microwave Remote Sensing. In *Satellite Soil Moisture Retrieval* (pp. 159-168). Elsevier.

Gupta, S., Kar, A. K., Baabdullah, A., & Al-Khowaiter, W. A. (2018). Big data with cognitive computing: A review for the future. *International Journal of Information Management*, *42*, 78-89.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM).* Sage publications.

Hamel, P., & Bryant, B. P. (2017). Uncertainty assessment in ecosystem services analyses: seven challenges and practical responses. *Ecosystem Services*, *24*, 1-15.

Hamida, A. B., Benoit, A., Lambert, P., & Amar, C. B. (2018). 3-D deep learning approach for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(8), 4420-4434.

Han, P., Zhang, W., Wang, G., Sun, W., & Huang, Y. (2016). Changes in soil organic carbon in croplands subjected to fertilizer management: a global meta-analysis. *Scientific reports*, *6*(1), 1-13.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., & Bauer-Marschallinger, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One, 12*(2), e0169748.

Heuvelink, G. B., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., & Olmedo, G. F. (2021). Machine learning in space and time for modelling soil organic carbon change. *European journal of soil science*, *72*(4), 1607-1623.

Hijbeek, R., Loon, M. P. v., & Ittersum, M. K. v. (2019). Fertiliser use and soil carbon sequestration: trade-offs and opportunities. *CCAFS Working Paper*.

Hijmans, R. J., & Graham, C. H. (2006). The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, *12*(12), 2272-2281.

Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade* (pp. 599-619). Springer.

Hively, W. D., Lamb, B. T., Daughtry, C. S., Shermeyer, J., McCarty, G. W., & Quemada, M. (2018). Mapping crop residue and tillage intensity using WorldView-3 satellite shortwave infrared residue indices. *Remote Sensing*, *10*(10), 1657.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hoffmann, U., Hoffmann, T., Johnson, E., & Kuhn, N. J. (2014). Assessment of variability and uncertainty of soil organic carbon in a mountainous boreal forest (Canadian Rocky Mountains, Alberta). *Catena*, *113*, 107-121.

Hong, Y., Chen, S., Zhang, Y., Chen, Y., Yu, L., Liu, Y., Liu, Y., Cheng, H., & Liu, Y. (2018). Rapid identification of soil organic matter level via visible and near-infrared spectroscopy: Effects of two-dimensional correlation coefficient and extreme learning machine. *Science of The Total Environment*, *644*, 1232-1243.

Houghton, E. (1996). *Climate change 1995: The science of climate change: contribution of working group I to the second assessment report of the Intergovernmental Panel on Climate Change* (Vol. 2). Cambridge University Press.

Huang, W., Xiao, L., Wei, Z., Liu, H., & Tang, S. (2015). A new pan-sharpening method with deep neural networks. *IEEE Geoscience and Remote Sensing Letters*, *12*(5), 1037-1041.

Huete, A., Justice, C., & Van Leeuwen, W. (1999). MODIS vegetation index (MOD13). *Algorithm theoretical basis document*, *3*, 213.

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, *25*(3), 295-309.

IPCC. (2016). Intergovernmental Panel on Climate Change (IPCC) (2016). http://www.ipcc.ch/ (accessed January 2016).

IPCC. (2021). *IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. In: MASSON-DELMOTTE, V., P. ZHAI, A. PIRANI, S.L. CONNORS, C. PÉAN, S. BERGER, N. CAUD, Y. CHEN, L. GOLDFARB, M.I. GOMIS, M. HUANG,* *K. LEITZELL, E. LONNOY, J.B.R. MATTHEWS, T.K. MAYCOCK, T. WATERFIELD, O. YELEKÇI, R. YU, AND B. ZHOU (ed.).*

Jäättelä, R. South Africa–Improved Understan-ding for Renewable Energy Potential. *Identifying Possibilities and Building Networks for Renewable Energy in Nigeria, Kenya and South Africa: Connect Project Experiences*, 46.

Jaber, S. M., Lant, C. L., & Al-Qinna, M. I. (2011). Estimating spatial variations in soil organic carbon using satellite hyperspectral data and map algebra. *International Journal of Remote Sensing*, *32*(18), 5077-5103.

Jackson, R. B., Schenk, H., Jobbagy, E., Canadell, J., Colello, G., Dickinson, R., Field, C., Friedlingstein, P., Heimann, M., & Hibbard, K. (2000). Belowground consequences of vegetation change and their treatment in models. *Ecological Applications*, *10*(2), 470-483.

Jamalabad, M., & Abkar, A. (2004). Forest canopy density monitoring, using satellite images. ISPRS Congress, Istanbul. In.

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Janik, L., Forrester, S., & Rawson, A. (2009). The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. *Chemometrics and Intelligent Laboratory Systems*, *97*(2), 179-188.

Jeong, G., Oeverdieck, H., Park, S. J., Huwe, B., & Ließ, M. (2017). Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. *Catena*, *154*, 73-84.

Jiao, S., Li, J., Li, Y., Xu, Z., Kong, B., Li, Y., & Shen, Y. (2020). Variation of soil organic carbon and physical properties in relation to land uses in the Yellow River Delta, China. *Scientific reports*, *10*(1), 1-12.

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.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Jost, E., Schönhart, M., Skalský, R., Balkovič, J., Schmid, E., & Mitter, H. (2021). Dynamic soil functions assessment employing land use and climate scenarios at regional scale. *Journal of environmental management*, *287*, 112318.

Kamaraj, M., & Rangarajan, S. (2022). Predicting the future land use and land cover changes for Bhavani basin, Tamil Nadu, India, using QGIS MOLUSCE plugin. *Environmental Science and Pollution Research*, 1-12.

Kasai, M., Marutani, T., Reid, L. M., & Trustrum, N. A. (2001). Estimation of temporally averaged sediment delivery ratio using aggradational terraces in headwater catchments of the Waipaoa River, North Island, New Zealand. *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group*, *26*(1), 1-16.

Kenye, A., Sahoo, U. K., Singh, S. L., & Gogoi, A. (2019). Soil organic carbon stock of different land uses of Mizoram, Northeast India. *AIMS Geosciences*, *5*(1), 25-40.

Keskin, H., Grunwald, S., & Harris, W. G. (2019). Digital mapping of soil carbon fractions with machine learning. *Geoderma*, *339*, 40-58.

Kesselmeier, J., Ciccioli, P., Kuhn, U., Stefani, P., Biesenthal, T., Rottenberger, S., Wolf, A., Vitullo, M., Valentini, R., & Nobre, A. (2002). Volatile organic compound emissions in relation to plant carbon fixation and the terrestrial carbon budget. *Global Biogeochemical Cycles*, *16*(4), 73-71-73-79.

Khapayi, M., & Celliers, P. (2016). Factors limiting and preventing emerging farmers to progress to commercial agricultural farming in the King William's Town area of the Eastern Cape Province, South Africa. *South African Journal of Agricultural Extension*, *44*(1), 25-41.

Koch, B. (2010). Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment. *ISPRS J. Photogramm. Remote Sens.*, *65*(6), 581-590.

Köchy, M., Hiederer, R., & Freibauer, A. (2015). Global distribution of soil organic carbon–Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world. *Soil*, *1*(1), 351-365.

Kokhanovsky, A., Lamare, M., Danne, O., Brockmann, C., Dumont, M., Picard, G., Arnaud, L., Favier, V., Jourdain, B., & Le Meur, E. (2019). Retrieval of snow properties from the Sentinel-3 Ocean and Land Colour Instrument. *Remote Sensing*, *11*(19), 2280.

Koteen, L. E., Baldocchi, D. D., & Harte, J. (2011). Invasion of non-native grasses causes a drop in soil carbon storage in California grasslands. *Environmental Research Letters*, *6*(4), 044001.

Kruse, F. A., Baugh, W. M. & Perry, S. L. (2015). Validation Of Digitalglobe Worldview-3 Earth Imaging Satellite Shortwave Infrared Bands For Mineral Mapping. *Journal Of Applied Remote Sensing, 9, 096044-096044.*

Kuang, B., Tekin, Y., & Mouazen, A. M. (2015). Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil and Tillage Research*, *146*, 243-252.

Kumar, L., & Mutanga, O. (2018). Google Earth Engine applications since inception: Usage, trends, and potential. *Remote Sensing*, *10*(10), 1509.

Kumar, P., Pandey, P. C., Singh, B., Katiyar, S., Mandal, V., Rani, M., Tomar, V., & Patairiya, S. (2016). Estimation of accumulated soil organic carbon stock in tropical forest using geospatial strategy. *The Egyptian Journal of Remote Sensing and Space Science*, *19*(1), 109-123.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J Stat Softw*, *36*(11), 1-13.

Laganiere, J. m., Paré, D., Bergeron, Y., Chen, H. Y., Brassard, B. W., & Cavard, X. (2013). Stability of soil carbon stocks varies with forest composition in the Canadian boreal biome. *Ecosystems*, *16*(5), 852-865.

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, *304*(5677), 1623-1627.

Lal, R. (2008). Promise and limitations of soils to minimize climate change. *Journal of soil and water conservation*, *63*(4), 113A-118A.

Lal, R. (2009). Sequestering carbon in soils of arid ecosystems. *Land Degradation & Development*, *20*(4), 441-454.

Lal, R. (2015). Restoring soil quality to mitigate soil degradation. *Sustainability*, *7*(5), 5875-5895.

Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma, 352*, 395-413.

Lang, M., McCarty, G., Oesterling, R., & Yeo, I.-Y. (2013). Topographic metrics for improved mapping of forested wetlands. *Wetlands, 33*(1), 141-155.

Laurin, G. V., Chen, Q., Lindsell, J. A., Coomes, D. A., Del Frate, F., Guerriero, L., Pirotti, F., & Valentini, R. (2014). Aboveground biomass estimation in an African tropical forest with lidar and hyperspectral data. *ISPRS J. Photogramm. Remote Sens.*, *89*, 49-58.

Leblois, A., Damette, O., & Wolfersberger, J. (2017). What has driven deforestation in developing countries since the 2000s? Evidence from new remote-sensing data. *World Development, 92*, 82-102.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Leone, A. P., Leone, N., & Rampone, S. (2013). An application of vis-NIR reflectance spectroscopy and artificial neural networks to the prediction of soil organic carbon content in southern Italy. *Fresenius Environmental Bulletin, 22*(4B), 1230-1238.

Levine, E., Kimes, D., Fifer, S., Nelson, R., & Lal, R. (2000). Evaluating tropical soil properties with pedon data, satellite imagery, and neural networks. *Global climate change and tropical ecosystems*, 365-374.

Li, J., & Roy, D. P. (2017). A global analysis of Sentinel-2A, Sentinel-2B and Landsat-8 data revisit intervals and implications for terrestrial monitoring. *Remote Sensing, 9*(9), 902.

Li, Q.-Q., Wang, C.-Q., Zhang, W.-J., Yu, Y., Li, B., Yang, J., Bai, G.-C., & Cai, Y. (2013). Prediction of soil nutrients spatial distribution based on neural network model combined with goestatistics. *Ying yong sheng tai xue bao= The journal of applied ecology, 24*(2), 459-466.

Li, Q.-q., Yue, T.-x., Wang, C.-q., Zhang, W.-j., Yu, Y., Li, B., Yang, J., & Bai, G.-c. (2013). Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach. *Catena*, *104*, 210-218.

Li, Q.-Q., Zhang, X., Wang, C.-Q., Li, B., Gao, X.-S., Yuan, D.-G., & Luo, Y.-L. (2016). Spatial prediction of soil nutrient in a hilly area using artificial neural network model combined with kriging. *Archives of Agronomy and Soil Science, 62*(11), 1541-1553.

Li, Q., Fang, H., Sun, L., & Cai, Q. (2014). Using the 137Cs technique to study the effect of soil redistribution on soil organic carbon and total nitrogen stocks in an agricultural catchment of Northeast China. *Land Degradation & Development*, *25*(4), 350-359.

Li, X., Ding, J., Liu, J., Ge, X., & Zhang, J. (2021). Digital Mapping of Soil Organic Carbon Using Sentinel Series Data: A Case Study of the Ebinur Lake Watershed in Xinjiang. *Remote Sensing*, *13*(4), 769.

Li, X., McCarty, G. W., Karlen, D. L., & Cambardella, C. A. (2018). Topographic metric predictions of soil redistribution and organic carbon in Iowa cropland fields. *Catena, 160*, 222-232.

Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(6), e1264.

Liao, C., Peng, R., Luo, Y., Zhou, X., Wu, X., Fang, C., Chen, J., & Li, B. (2008). Altered ecosystem carbon and nitrogen cycles by plant invasion: a meta-analysis. *New Phytologist*, *177*(3), 706-714.

Ließ, M., Schmidt, J., & Glaser, B. (2016). Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches. *PLoS One*, *11*(4), e0153673.

Lin, C., Zhu, A.-X., Wang, Z., Wang, X., & Ma, R. (2020). The refined spatiotemporal representation of soil organic matter based on remote images fusion of Sentinel-2 and Sentinel-3. *International Journal of Applied Earth Observation and Geoinformation*, *89*, 102094.

Lin, S., Liu, X., Fang, J., & Xu, Z. (2014). Is extreme learning machine feasible? A theoretical assessment (Part II). *IEEE Transactions on Neural Networks and Learning Systems*, *26*(1), 21-34.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60-88.

Little, I. T., Hockey, P. A., & Jansen, R. (2015). Impacts of fire and grazing management on South Africa's moist highland grasslands: A case study of the Steenkampsberg Plateau, Mpumalanga, South Africa. *Bothalia-African Biodiversity & Conservation*, *45*(1), 1-15.

Liu, D., Wang, Z., Zhang, B., Song, K., Li, X., Li, J., Li, F., & Duan, H. (2006). Spatial distribution of soil organic carbon and analysis of related factors in croplands of the black soil region, Northeast China. *Agriculture, Ecosystems & Environment*, *113*(1-4), 73-81.

Liu, F., Zhang, G.-L., Sun, Y.-J., Zhao, Y.-G., & Li, D.-C. (2013). Mapping the Three-Dimensional Distribution of Soil Organic Matter across a Subtropical Hilly Landscape. *Soil Science Society of America Journal*, *77*(4), 1241-1253.

Liu, W., Zhu, M., Li, Y., Zhang, J., Yang, L., & Zhang, C. (2021). Assessing Soil Organic Carbon Stock Dynamics under Future Climate Change Scenarios in the Middle Qilian Mountains. *Forests*, *12*(12), 1698.

Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., & Wang, X. (2018). Deep learning for pixel-level image fusion: Recent advances and future prospects. *Information Fusion*, *42*, 158-173.

Liu, Z., Shao, M. a., & Wang, Y. (2011). Effect of environmental factors on regional soil organic carbon stocks across the Loess Plateau region, China. *Agriculture, Ecosystems & Environment*, *142*(3-4), 184-194.

Lugato, E., Panagos, P., Bampa, F., Jones, A., & Montanarella, L. (2014). A new baseline of organic carbon stock in European agricultural soils using a modelling approach. *Global Change Biology*, *20*(1), 313-326.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 166-177.

Maddison, C. J., Mnih, A., & Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Madileng, N. P., Mutanga, O., Dube, T., & Odebiri, O. (2020). Mapping the spatial distribution of Lantana camara using high-resolution SPOT 6 data, in Mpumalanga communal areas, South Africa. *Transactions of the Royal Society of South Africa*, *75*(3), 239-244.

Makambo, K. A., & Kisaka, E. S. (2017). The Effect of Carbon Sequestration on Farmers' Income: A Case Study of Kenya Agricultural Carbon Project. *African Development Finance Journal (ADFJ)*, *1*(2).

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International journal of forecasting*, *9*(4), 527-529.

Margenot, A., O'Neill, T., Sommer, R., & Akella, V. (2020). Predicting soil permanganate oxidizable carbon (POXC) by coupling DRIFT spectroscopy and artificial neural networks (ANN). *Computers and Electronics in Agriculture*, *168*, 105098.

Masemola, C., Cho, M. A., & Ramoelo, A. (2020). Towards a semi-automated mapping of Australia native invasive alien Acacia trees using Sentinel-2 and radiative transfer models in South Africa. *ISPRS Journal of Photogrammetry and Remote Sensing*, *166*, 153-168.

Masemola, C. R., & Cho, M. A. (2019). Similarities of spectral bands from intact fresh and dry leaves spectra for estimating leaf nitrogen concentration using model population analysis framework.

Matsushita, B., Yang, W., Chen, J., Onda, Y., & Qiu, G. (2007). Sensitivity of the enhanced vegetation index (EVI) and normalized difference vegetation index (NDVI) to topographic effects: a case study in high-density cypress forest. *Sensors*, *7*(11), 2636-2651.

Melillo, J. M., Frey, S. D., DeAngelis, K. M., Werner, W. J., Bernard, M. J., Bowles, F. P., Pold, G., Knorr, M. A., & Grandy, A. S. (2017). Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. *Science*, *358*(6359), 101-105.

Meng, X., Bao, Y., Liu, J., Liu, H., Zhang, X., Zhang, Y., Wang, P., Tang, H., & Kong, F. (2020). Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. *International Journal of Applied Earth Observation and Geoinformation*, *89*, 102111.

Mills, A., & Fey, M. (2003). Declining soil quality in South Africa: effects of land use on soil organic matter and surface crusting. *South African Journal of Science*, *99*(9), 429-436.

Mills, A., & Fey, M. (2004). Soil carbon and nitrogen in five contrasting biomes of South Africa exposed to different land uses. *South African Journal of Plant and Soil*, *21*(2), 94-103.

Mills, A. J., & Cowling, R. M. (2014). How fast can carbon be sequestered when restoring degraded subtropical thicket? *Restoration Ecology*, *22*(5), 571-573.

Mills, A. J., Vyver, M. V. d., Gordon, I. J., Patwardhan, A., Marais, C., Blignaut, J., Sigwela, A., & Kgope, B. (2015). Prescribing innovation within a large-scale restoration programme in degraded subtropical thicket in South Africa. *Forests*, *6*(11), 4328-4348.

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., & Das, B. S. (2017). Soil carbon 4 per mille. *Geoderma*, *292*, 59-86.

Minasny, B., Setiawan, B. I., Arif, C., Saptomo, S. K., & Chadirin, Y. (2016). Digital mapping for cost-effective and accurate prediction of the depth and carbon stocks in Indonesian peatlands. *Geoderma*, *272*, 20-31.

Minh, D. H. T., Ienco, D., Gaetano, R., Lalande, N., Ndikumana, E., Osman, F., & Maurel, P. (2018). Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR Sentinel-1. *IEEE Geoscience and Remote Sensing Letters*, *15*(3), 464-468.

Mirzaee, S., Ghorbani-Dashtaki, S., Mohammadi, J., Asadi, H., & Asadzadeh, F. (2016). Spatial variability of soil organic matter using remote sensing data. *Catena*, *145*, 118-127.

Mishra, U., & Mapa, R. B. (2019). National soil organic carbon estimates can improve global estimates. *Geoderma*, *337*, 55-64.

Mngadi, M., Odindi, J., Peerbhay, K., & Mutanga, O. (2019). Examining the effectiveness of Sentinel-1 and 2 imagery for commercial forest species mapping. *Geocarto International*, 1-12.

Mngadi, M., Odindi, J., Peerbhay, K., Mutanga, O., & Sibanda, M. (2020). Testing the utility of multivariate techniques in mapping commercial forest species using freely available Landsat 8 Operational Land Imager (OLI). *Journal of Forest Research*, 1-4.

Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., & Mukhopadhyay, A. (2017). Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data. *The Egyptian Journal of Remote Sensing and Space Science*, *20*(1), 61-70.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247-259.

Moyo, H., Slotow, R., Rouget, M., Mugwedi, L., Douwes, E., Tsvuura, Z., & Tshabalala, T. (2021). Adaptive management in restoration initiatives: Lessons learned from some of South Africa's projects. *South African Journal of Botany*, *139*, 352-361.

Muchena, R. (2017). Estimating Soil Carbon Stocks in a Dry Miombo Ecosystem Using Remote Sensing. *Forest Res*, *6*(198), 2.

Mucina, L., & Rutherford, M. C. (2006). *The vegetation of South Africa, Lesotho and Swaziland*. South African National Biodiversity Institute.

Mujinya, B., Mees, F., Erens, H., Dumon, M., Baert, G., Boeckx, P., Ngongo, M., & Van Ranst, E. (2013). Clay composition and properties in termite mounds of the Lubumbashi area, DR Congo. *Geoderma*, *192*, 304-315.

Mutanga, O., Adam, E., & Cho, M. A. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *Int J Appl Earth Obs*, *18*, 399-406. https://doi.org/10.1016/j.jag.2012.03.012

Mutanga, O., & Ismail, R. (2015). Remote sensing bio-control damage on aquatic invasive alien plant species. *SAJG*, *4*(4), 464-485.

Mutanga, O., & Skidmore, A. K. (2004). Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing*, *25*(19), 3999-4014.

Mzinyane, T., van Aardt, J., & Gebreslasie, M. T. (2015). Soil carbon estimation from eucalyptus grandis using canopy spectra. *SAJG*, *4*(4), 548-561.

Nabiollahi, K., Eskandari, S., Taghizadeh-Mehrjardi, R., Kerry, R., & Triantafilis, J. (2019). Assessing soil organic carbon stocks under land-use change scenarios using random forest models. *Carbon Management*, *10*(1), 63-77.

Naidoo, R., & Fisher, B. (2020). Reset sustainable development goals for a pandemic world. In: Nature Publishing Group.

Ndalowa, D. (2014). *Evaluation of carbon accounting models for plantation forestry in South Africa* Stellenbosch: Stellenbosch University].

Newell, R. G., Prest, B. C., & Sexton, S. E. (2021). The GDP-temperature relationship: implications for climate change damages. *Journal of Environmental Economics and Management*, *108*, 102445.

Novoa, J., Fredes, J., Poblete, V., & Yoma, N. B. (2018). Uncertainty weighting and propagation in DNN–HMM-based speech recognition. *Computer Speech & Language*, *47*, 30-46.

O'BRIEN, S. L., Jastrow, J. D., Grimley, D. A., & GONZALEZ-MELER, M. A. (2010). Moisture and vegetation controls on decadal-scale accrual of soil organic carbon and total nitrogen in restored grasslands. *Global Change Biology*, *16*(9), 2573-2588.

Odebiri, O., Mutanga, O., & Odindi, J. (2022a). Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data. *Geoderma*, *411*, 115695.

Odebiri, O., Mutanga, O., Odindi, J., & Naicker, R. (2022b). Modelling soil organic carbon stock distribution across different land-uses in South Africa: A remote sensing and deep learning approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, *188*, 351-362.

Odebiri, O., Mutanga, O., Odindi, J., Naicker, R., Masemola, C., & Sibanda, M. (2021b). Deep learning approaches in remote sensing of soil organic carbon: a review of utility, challenges, and prospects. *Environmental monitoring and assessment*, *193*(12), 1-18.

Odebiri, O., Mutanga, O., Odindi, J., Peerbhay, K., & Dovey, S. (2020a). Predicting soil organic carbon stocks under commercial forest plantations in KwaZulu-Natal province, South Africa using remotely sensed data. *GIScience & Remote Sensing*, 1-14.

Odebiri, O., Mutanga, O., Odindi, J., Peerbhay, K., Dovey, S., & Ismail, R. (2020b). Estimating soil organic carbon stocks under commercial forestry using topo-climate variables in KwaZulu-Natal, South Africa. *South African Journal of Science*, *116*(3-4), 1-8.

Odebiri, O., Odindi, J., & Mutanga, O. (2021a). Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review. *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102389.

Odindi, J., Bangamwabo, V., & Mutanga, O. (2015). Assessing theValue ofUrbanGreen Spaces inMitigatingMulti-SeasonalUrban Heat usingMODISLand SurfaceTemperature (LST) andLandsat 8 data. *International Journal of Environmental Research*, *9*(1), 9-18.

Odindi, J., Mutanga, O., Rouget, M., & Hlanguza, N. (2016). Mapping alien and indigenous vegetation in the KwaZulu-Natal Sandstone Sourveld using remotely sensed data. *Bothalia*, *46*(2), 1-9.

Ogbodo, J. A., Wasige, E. J., Shuaibu, S. M., Dube, T., & Anarah, S. E. (2019). Remote Sensing of Droughts Impacts on Maize Prices Using SPOT-VGT Derived Vegetation Index. In *Climate Change-Resilient Agriculture and Agroforestry* (pp. 235-255). Springer.

Olsson, L., & Ardö, J. (2002). Soil carbon sequestration in degraded semiarid agro-ecosystems—perils and potentials. *AMBIO: A Journal of the Human Environment*, *31*(6), 471-477.

Ontl, T., & Iversen, C. (2017). *SPRUCE S1 bog areal coverage of hummock and hollow microtopography assessed along three transects in the S1 bog*.

Ou, G., Tan, S., Zhou, M., Lu, S., Tao, Y., Zhang, Z., Zhang, L., Yan, D., Guan, X., & Wu, G. (2017). An interval chance-constrained fuzzy modeling approach for supporting land-use planning and eco-environment planning at a watershed level. *Journal of environmental management*, *204*, 651-666.

Owusu, S., Yigini, Y., Olmedo, G. F., & Omuto, C. T. (2020). Spatial prediction of soil organic carbon stocks in Ghana using legacy data. *Geoderma*, *360*, 114008.

Padarian, J., Minasny, B., & McBratney, A. (2019a). Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma*, *340*, 279-288.

Padarian, J., Minasny, B., & McBratney, A. (2019b). Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional*, *16*, e00198.

Padarian, J., Minasny, B., & McBratney, A. B. (2019c). Using deep learning for digital soil mapping. *Soil*, *5*(1), 79-89.

Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *Soil*, *6*(1), 35-52.

Padarian, J., Minasny, B., McBratney, A. B., & Smith, P. (2021). Additional soil organic carbon storage potential in global croplands. *SOIL Discussions*, 1-15.

Palmer, A. R., & Ainslie, A. M. (2005). Grasslands of South Africa. *Grasslands of the World*, *34*, 77.

Pang, S., & Yang, X. (2016). Deep convolutional extreme learning machine and its application in handwritten digit classification. *Computational Intelligence and Neuroscience*, *2016*.

Paterson, D. G. (2014). *Soil information for proposed upington solar park, northern cape*

Pearson, T. R., Brown, S. L., & Birdsey, R. A. (2007). Measurement guidelines for the sequestration of forest carbon. *Gen. Tech. Rep. NRS-18. Newtown Square, PA: USDA, Forest Service, Northern Research Station. 42 p.*, *18*.

Peh, K. S.-H., Balmford, A., Birch, J. C., Brown, C., Butchart, S. H., Daley, J., Dawson, J., Gray, G., Hughes, F. M., & Mendes, S. (2015). Potential impact of invasive alien species on ecosystem services provided by a tropical forested ecosystem: a case study from Montserrat. *Biological Invasions*, *17*(1), 461-475.

Pentoś, K. (2016). The methods of extracting the contribution of variables in artificial neural network models–Comparison of inherent instability. *Computers and Electronics in Agriculture*, *127*, 141-146.

Peri, P. L. (2011). Carbon storage in cold temperate ecosystems in Southern Patagonia, Argentina. *Biomass and remote sensing of biomass*, 213-226.

Phachomphon, K., Dlamini, P., & Chaplot, V. (2010). Estimating carbon stocks at a regional level using soil information and easily accessible auxiliary variables. *Geoderma*, *155*(3-4), 372-380.

Pudełko, A., & Chodak, M. (2020). Estimation of total nitrogen and organic carbon contents in mine soils with NIR reflectance spectroscopy and various chemometric methods. *Geoderma*, *368*, 114306.

Qi, J., Chehbouni, A., Huete, A., Kerr, Y., & Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote Sensing of Environment*, *48*(2), 119-126.

Ramesh, T., Bolan, N. S., Kirkham, M. B., Wijesekara, H., Kanchikerimath, M., Rao, C. S., Sandeep, S., Rinklebe, J., Ok, Y. S., & Choudhury, B. U. (2019). Soil organic carbon dynamics: Impact of land use changes and management practices: A review. *Advances in agronomy*, *156*, 1-107.

Ramjee, S., & Gamal, A. E. (2019). Efficient wrapper feature selection using autoencoder and model based elimination. *arXiv preprint arXiv:1905.11592*.

Rasaei, Z., & Bogaert, P. (2019). Spatial filtering and Bayesian data fusion for mapping soil properties: A case study combining legacy and remotely sensed data in Iran. *Geoderma*, *344*, 50-62.

Ren, W., Banger, K., Tao, B., Yang, J., Huang, Y., & Tian, H. (2020). Global pattern and change of cropland soil organic carbon during 1901-2010: Roles of climate, atmospheric chemistry, land use and management. *Geography and Sustainability*, *1*(1), 59-69.

Reyna-Bowen, L., Fernandez-Rebollo, P., Fernández-Habas, J., & Gómez, J. A. (2020). The influence of tree and soil management on soil organic carbon stock and pools in dehesa systems. *Catena*, *190*, 104511.

Rezaei, S. A., & Gilkes, R. J. (2005). The effects of landscape attributes and plant community on soil physical properties in rangelands. *Geoderma*, *125*(1-2), 145-154.

Richardson, A. J., & Wiegand, C. (1977). Distinguishing vegetation from soil background information. *Photogrammetric engineering and remote sensing*, *43*(12), 1541-1552.

Richardson, H. J., Hill, D. J., Denesiuk, D. R., & Fraser, L. H. (2017). A comparison of geographic datasets and field measurements to model soil carbon using random forests and stepwise regressions (British Columbia, Canada). *GIScience & Remote Sensing*, *54*(4), 573-591.

Ritchie, J. C., McCarty, G. W., Venteris, E. R., & Kaspar, T. (2007). Soil and soil organic carbon redistribution on the landscape. *Geomorphology*, *89*(1-2), 163-171.

Roberts, D., Boon, R., Diederichs, N., Douwes, E., Govender, N., Mcinnes, A., Mclean, C., O'Donoghue, S., & Spires, M. (2012). Exploring ecosystem-based adaptation in Durban, South Africa:"learning-by-doing" at the local government coal face. *Environment and Urbanization*, *24*(1), 167-195.

Roberts, J. L. (2021). Climate Change and Heatwaves. In *Shaping the Future of Small Islands* (pp. 233-248). Springer.

Rodriguez, F., Maire, E., Courjault-Radé, P., & Darrozes, J. (2002). The Black Top Hat function applied to a DEM: A tool to estimate recent incision in a mountainous watershed (Estibère Watershed, Central Pyrenees). *Geophysical research letters*, *29*(6).

Romero, A., Gatta, C., & Camps-Valls, G. (2015). Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(3), 1349-1362.

Rossel, R. V., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, *158*(1-2), 46-54.

Rouget, M., Naicker, R., & Mutanga, O. (2016). Assessing habitat fragmentation of the KwaZulu-Natal Sandstone Sourveld, a threatened ecosystem. *Bothalia-African Biodiversity & Conservation*, *46*(2), 1-10.

Roujean, J.-L., & Breon, F.-M. (1995). Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment*, *51*(3), 375-384.

Rouse Jr, J. W., Haas, R., Schell, J., & Deering, D. (1974). Monitoring vegetation systems in the Great Plains with ERTS.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536.

Rusere, F., Crespo, O., Mkuhlani, S., & Dicks, L. V. (2019). Developing pathways to improve smallholder agricultural productivity through ecological intensification technologies in semi-arid Limpopo, South Africa. *African Journal of Science, Technology, Innovation and Development*, *11*(5), 543-553.

Rutherford, M. C., Mucina, L., & Powrie, L. W. (2006). Biomes and bioregions of southern Africa. *The vegetation of South Africa, Lesotho and Swaziland, 19,* 30-51.

Sahoo, U. K., Singh, S. L., Gogoi, A., Kenye, A., & Sahoo, S. S. (2019). Active and passive soil organic carbon pools as affected by different land use types in Mizoram, Northeast India. *PLoS One, 14*(7).

Sainepo, B. M., Gachene, C. K., & Karuma, A. (2018). Assessment of soil organic carbon fractions and carbon management index under different land use types in Olesharo Catchment, Narok County, Kenya. *Carbon balance and management, 13*(1), 1-9.

Samek, D., & Dostal, P. (2009). Artificial neural network with radial basis function in model predictive control of chemical reactor. *Mechanics/AGH University of Science and Technology, 28*(3), 91-95.

Sanderman, J., Hengl, T., & Fiske, G. J. (2017). Soil carbon debt of 12,000 years of human land use. *Proceedings of the National Academy of Sciences, 114*(36), 9575-9580.

SANLC. (2020). South African National Land-Cover 2020 Report & Accuracy Assessment. *Department of Environment, Forestry & Fisheries, Pretoria, South Africa (2020)*.

Schimel, D., Pavlick, R., Fisher, J. B., Asner, G. P., Saatchi, S., Townsend, P., Miller, C., Frankenberg, C., Hibbard, K., & Cox, P. (2015). Observing terrestrial ecosystems and the carbon cycle from space. *Global Change Biology, 21*(5), 1762-1776.

Schindlbacher, A., Wunderlich, S., Borken, W., Kitzler, B., Zechmeister-Boltenstern, S., & Jandl, R. (2012). Soil respiration under climate change: prolonged summer drought offsets soil warming effects. *Global Change Biology, 18*(7), 2270-2279.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Schulze, R. E., & Schütte, S. (2020). Mapping soil organic carbon at a terrain unit resolution across South Africa. *Geoderma, 373*, 114447.

Schwanghart, W., & Jarmer, T. (2011). Linking spatial patterns of soil organic carbon to topography—A case study from south-eastern Spain. *Geomorphology, 126*(1-2), 252-263.

Scott, D. F., & Lesch, W. (1997). Streamflow responses to afforestation with Eucalyptus grandis and Pinus patula and to felling in the Mokobulaan experimental catchments, South Africa. *Journal of Hydrology, 199*(3-4), 360-377.

Seijmonsbergen, A. C., Hengl, T., & Anders, N. S. (2011). Semi-automated identification and extraction of geomorphological features using digital elevation data. In *Developments in earth surface processes* (Vol. 15, pp. 297-335). Elsevier.

Shao, Z., Zhang, L., & Wang, L. (2017). Stacked sparse autoencoder modeling using the synergy of airborne LiDAR and satellite optical and SAR data to map forest above-ground biomass. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(12), 5569-5582.

Sharma, G., Sharma, L., & Sharma, K. (2019). Assessment of land use change and its effect on soil carbon stock using multitemporal satellite data in semiarid region of Rajasthan, India. *Ecological Processes, 8*(1), 1-17.

Shelukindo, H. B., Semu, E., Msanya, B., Munishi, P. K., Maliondo, S., & Singh, B. R. (2014). Potential of carbon storage in major soil types of the Miombo woodland ecosystem, Tanzania: A review.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., & Fang, Z. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences (Online), 22*(11).

Shen, H., Li, T., Yuan, Q., & Zhang, L. (2018). Estimating regional ground-level PM2. 5 directly from satellite top-of-atmosphere reflectance using deep belief networks. *Journal of Geophysical Research: Atmospheres*, *123*(24), 13,875-813,886.

Shukla, M., Lal, R., & Ebinger, M. (2006). Determining soil quality indicators by factor analysis. *Soil Till Res*, *87*(2), 194-204.

Sibanda, M., Mutanga, O., Rouget, M., & Odindi, J. (2015). Exploring the potential of in situ hyperspectral data and multivariate techniques in discriminating different fertilizer treatments in grasslands. *Journal of Applied Remote Sensing*, *9*(1), 096033.

Sibanda, M., Onisimo, M., Dube, T., & Mabhaudhi, T. (2021). Quantitative assessment of grassland foliar moisture parameters as an inference on rangeland condition in the mesic rangelands of southern Africa. *International Journal of Remote Sensing*, *42*(4), 1474-1491.

Siewert, M. B. (2018). High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: a case study in a sub-Arctic peatland environment. *Biogeosciences*, *15*(6), 1663-1682.

Singh, S., & Kasana, S. S. (2019). Estimation of soil properties from the EU spectral library using long short-term memory networks. *Geoderma Regional*, *18*, e00233.

Singh, S., Pandey, C., Sidhu, G., Sarkar, D., & Sagar, R. (2011). Concentration and stock of carbon in the soils affected by land uses and climates in the western Himalaya, India. *Catena*, *87*(1), 78-89.

Sirsat, M., Cernadas, E., Fernández-Delgado, M., & Barro, S. (2018). Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods. *Computers and Electronics in Agriculture*, *154*, 120-133.

Somarathna, P., Minasny, B., & Malone, B. P. (2017). More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon. *Soil Science Society of America Journal*, *81*(6), 1413-1426.

Song, Y.-Q., Yang, L.-A., Li, B., Hu, Y.-M., Wang, A.-L., Zhou, W., Cui, X.-S., & Liu, Y.-L. (2017). Spatial prediction of soil organic matter using a hybrid geostatistical model of an extreme learning machine and ordinary kriging. *Sustainability*, *9*(5), 754.

Sulman, B. N., Moore, J. A., Abramoff, R., Averill, C., Kivlin, S., Georgiou, K., Sridhar, B., Hartman, M. D., Wang, G., & Wieder, W. R. (2018). Multiple models and experiments underscore large uncertainty in soil carbon dynamics. *Biogeochemistry*, *141*(2), 109-123.

Swanepoel, C., Van der Laan, M., Weepener, H., Du Preez, C., & Annandale, J. G. (2016). Review and meta-analysis of organic matter in cultivated soils in southern Africa. *Nutrient cycling in agroecosystems*, *104*(2), 107-123.

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., & Scholten, T. (2020). Improving the Spatial Prediction of Soil Organic Carbon Content in Two Contrasting Climatic Regions by Stacking Machine Learning Models and Rescanning Covariate Space. *Remote Sensing*, *12*(7), 1095.

Tekin, Y., Tümsavas, Z., & Mouazen, A. M. (2014). Comparing the artificial neural network with parcial least squares for prediction of soil organic carbon and pH at different moisture content levels using visible and near-infrared spectroscopy. *Revista Brasileira de Ciência do Solo, 38*(6), 1794-1804.

Tiefenbacher, A., Sandén, T., Haslmayr, H.-P., Miloczki, J., Wenzel, W., & Spiegel, H. (2021). Optimizing Carbon Sequestration in Croplands: A Synthesis. *Agronomy, 11*(5), 882.

Troch, P., Van Loon, E., & Hilberts, A. (2002). Analytical solutions to a hillslope-storage kinematic wave equation for subsurface flow. *Advances in Water Resources*, *25*(6), 637-649.

Trumper, K. (2009). *The natural fix?: the role of ecosystems in climate mitigation: a UNEP rapid response assessment*. UNEP/Earthprint.

Turpie, J. K., Marais, C., & Blignaut, J. N. (2008). The working for water programme: Evolution of a payments for ecosystem services mechanism that addresses both poverty and ecosystem service delivery in South Africa. *Ecological economics, 65*(4), 788-798.

Vaudour, E., Gomez, C., Fouad, Y., & Lagacherie, P. (2019). Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. *Remote Sensing of Environment*, *223*, 21-33.

Venter, Z. S., Hawkins, H.-J., Cramer, M. D., & Mills, A. J. (2021). Mapping soil organic carbon stocks and trends with satellite-driven high resolution maps over South Africa. *Science of The Total Environment*, *771*, 145384.

Venter, Z. S., Hawkins, H. J., & Cramer, M. D. (2017). Implications of historical interactions between herbivory and fire for rangeland management in African savannas. *Ecosphere, 8*(10), e01946.

Wadoux, A. M.-C. (2019b). Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma*, *351*, 59-70.

Wadoux, A. M. J., Padarian, J., & Minasny, B. (2019a). Multi-source data integration for soil mapping using deep learning. *Soil*, *5*(1), 107-119.

Wan, Q., Zhu, G., Guo, H., Zhang, Y., Pan, H., Yong, L., & Ma, H. (2019). Influence of Vegetation Coverage and Climate Environment on Soil Organic Carbon in the Qilian Mountains. *Scientific reports*, *9*(1), 1-9.

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., & Li Liu, D. (2018). High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. *Science of The Total Environment*, *630*, 367-378.

Wang, G., Zhou, Y., Xu, X., Ruan, H., & Wang, J. (2013). Temperature sensitivity of soil organic carbon mineralization along an elevation gradient in the Wuyi Mountains, China. *PLoS One*, *8*(1), e53914.

Wang, H., Zhang, X., Wu, W., & Liu, H. (2021). Prediction of Soil Organic Carbon under Different Land Use Types Using Sentinel-1/-2 Data in a Small Watershed. *Remote Sensing*, *13*(7), 1229.

Wang, K., Qi, Y., Guo, W., Zhang, J., & Chang, Q. (2021). Retrieval and Mapping of Soil Organic Carbon Using Sentinel-2A Spectral Images from Bare Cropland in Autumn. *Remote Sensing*, *13*(6), 1072.

Wang, S., Xu, L., Zhuang, Q., & He, N. (2021). Investigating the spatio-temporal variability of soil organic carbon stocks in different ecosystems of China. *Science of The Total Environment*, *758*, 143644.

Wang, S., Zhuang, Q., Yang, Z., Yu, N., & Jin, X. (2019). Temporal and spatial changes of soil organic carbon stocks in the forest area of Northeastern China. *Forests*, *10*(11), 1023.

Wang, T., Zhang, H., Lin, H., & Fang, C. (2016). Textural–spectral feature-based species classification of mangroves in Mai Po Nature Reserve from Worldview-3 imagery. *Remote Sensing*, *8*(1), 24.

Wang, X., Wang, J., & Zhang, J. (2012). Comparisons of three methods for organic and inorganic carbon in calcareous soils of northwestern China. *PLoS One*, *7*(8), e44334.

Wang, Y., Zhang, Z., Feng, L., Du, Q., & Runge, T. (2020). Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States. *Remote Sensing*, *12*(8), 1232.

Ward, S. E., Smart, S. M., Quirk, H., Tallowin, J. R., Mortimer, S. R., Shiel, R. S., Wilby, A., & Bardgett, R. D. (2016). Legacy effects of grassland management on soil carbon to depth. *Global Change Biology*, *22*(8), 2929-2938.

Wei, S., Zhao, Z., Yang, Q., & Ding, X. (2021). A Two-Stage Approach to the Estimation of High-Resolution Soil Organic Carbon Storage with Good Extension Capability. *Land*, *10*(5), 517.

Weiss, N., Faucherre, S., Lampiris, N., & Wojcik, R. (2017). Elevation-based upscaling of organic carbon stocks in High-Arctic permafrost terrain: a storage and distribution assessment for Spitsbergen, Svalbard. *Polar Research*, *36*(1), 1400363.

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.

Wieder, W. R., Hartman, M. D., Sulman, B. N., Wang, Y. P., Koven, C. D., & Bonan, G. B. (2018). Carbon cycle confidence and uncertainty: Exploring variation among soil biogeochemical models. *Global Change Biology*, *24*(4), 1563-1579.

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 soil*, *340*(1-2), 7-24.

Wiesmeier, M., Poeplau, C., Sierra, C. A., Maier, H., Frühauf, C., Hübner, R., Kühnel, A., Spörlein, P., Geuß, U., & Hangen, E. (2016). Projected loss of soil organic carbon in temperate agricultural soils in the 21st century: effects of climate change and carbon input trends. *Scientific reports*, *6*(1), 1-17.

Wijewardane, N. K., Ge, Y., & Morgan, C. L. (2016). Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. *Geoderma*, *267*, 92-101.

Wijewardane, N. K., Ge, Y., Wills, S., & Libohova, Z. (2018). Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library. *Soil Science Society of America Journal*, *82*(3), 722-731.

Wittek, P. (2014). *Quantum machine learning: what quantum computing means to data mining*. Academic Press.

WMO. ((2021).). WMO Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (1970–2019). WMO-No. 1267. In: WMO Geneva.

Wolf, S., Eugster, W., Potvin, C., Turner, B. L., & Buchmann, N. (2011). Carbon sequestration potential of tropical pasture compared with afforestation in Panama. *Global Change Biology*, *17*(9), 2763-2780.

Woomer, P. L. (1993). The impact of cultivation on carbon fluxes in woody savannas of Southern Africa. *Water, Air, and Soil Pollution*, *70*(1), 403-412.

Wu, T., Luo, J., Dong, W., Sun, Y., Xia, L., & Zhang, X. (2019). Geo-object-based soil organic matter mapping using machine learning algorithms with multi-source geo-spatial data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(4), 1091-1106.

Wu, Y., Zhao, F., Liu, S., Wang, L., Qiu, L., Alexandrov, G., & Jothiprakash, V. (2018). Bioenergy production and environmental impacts. *Geoscience Letters*, *5*(1), 1-9.

Xiao, J., Chevallier, F., Gomez, C., Guanter, L., Hicke, J. A., Huete, A. R., Ichii, K., Ni, W., Pang, Y., & Rahman, A. F. (2019). Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. *Remote Sensing of Environment*, *233*, 111383.

Xu, S., Wang, M., & Shi, X. (2020). Hyperspectral imaging for high-resolution mapping of soil carbon fractions in intact paddy soil profiles with multivariate techniques and variable selection. *Geoderma*, *370*, 114358.

Xu, X., Du, C., Ma, F., Shen, Y., Wu, K., Liang, D., & Zhou, J. (2019). Detection of soil organic matter from laser-induced breakdown spectroscopy (LIBS) and mid-infrared spectroscopy (FTIR-ATR) coupled with multivariate techniques. *Geoderma*, *355*, 113905.

Xu, Z., Zhao, X., Guo, X., & Guo, J. (2019). Deep Learning Application for Predicting Soil Organic Matter Content by VIS-NIR Spectroscopy. *Computational Intelligence and Neuroscience*, *2019*.

Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., & Li, D.-C. (2016). Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecological Indicators*, *60*, 870-878.

Yao, M. K., Angui, P. K., Konaté, S., Tondoh, J. E., Tano, Y., Abbadie, L., & Benest, D. (2010). Effects of land use types on soil organic carbon and nitrogen dynamics in Mid-West Cote d'Ivoire. *European Journal of Scientific Research*, *40*(2), 211-222.

Yeasmin, S., Jahan, E., Molla, M., Islam, A., Anwar, M., Or Rashid, M., & Chungopast, S. (2020). Effect of land use on organic carbon storage potential of soils with contrasting native organic matter content. *International Journal of Agronomy*, *2020*.

Yigini, Y., & Panagos, P. (2016). Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. *Science of The Total Environment*, *557*, 838-850.

Yoo, K., Amundson, R., Heimsath, A. M., & Dietrich, W. E. (2006). Spatial patterns of soil organic carbon on hillslopes: Integrating geomorphic processes and the biological C cycle. *Geoderma*, *130*(1-2), 47-65.

Yu, H., Xie, T., Paszczynski, S., & Wilamowski, B. M. (2011). Advantages of radial basis function networks for dynamic system design. *IEEE Transactions on Industrial Electronics*, *58*(12), 5438-5450.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., & Wang, J. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, *241*, 111716.

Zaloumis, N. P., & Bond, W. J. (2016). Reforestation or conservation? The attributes of old growth grasslands in South Africa. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1703), 20150310.

Zhang, C., Ju, W., Chen, J. M., Li, D., Wang, X., Fan, W., Li, M., & Zan, M. (2014). Mapping forest stand age in China using remotely sensed forest height and observation data. *Journal of Geophysical Research: Biogeosciences, 119*(6), 1163-1179.

Zhang, C., Mishra, D. R., & Pennings, S. C. (2019). Mapping salt marsh soil properties using imaging spectroscopy. *ISPRS Journal of Photogrammetry and Remote Sensing, 148*, 221-234.

Zhang, L., Shao, Z., Liu, J., & Cheng, Q. (2019). Deep Learning based retrieval of forest aboveground biomass from combined LiDAR and Landsat 8 data. *Remote Sensing, 11*(12), 1459.

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine, 4*(2), 22-40.

Zhang, M., Zhang, M., Yang, H., Jin, Y., Zhang, X., & Liu, H. (2021). Mapping regional soil organic matter based on Sentinel-2A and MODIS imagery using machine learning algorithms and google earth engine. *Remote Sensing, 13*(15), 2934.

Zhang, Y., Guo, L., Chen, Y., Shi, T., Luo, M., Ju, Q., Zhang, H., & Wang, S. (2019). Prediction of Soil Organic Carbon based on Landsat 8 Monthly NDVI Data for the Jianghan Plain in Hubei Province, China. *Remote Sensing, 11*(14), 1683.

Zhang, Y., Zhao, Y., Shi, X., Lu, X., Yu, D., Wang, H., Sun, W., & Darilek, J. (2008). Variation of soil organic carbon estimates in mountain regions: a case study from Southwest China. *Geoderma, 146*(3-4), 449-456.

Zhao, F., Wu, Y., Hui, J., Sivakumar, B., Meng, X., & Liu, S. (2021). Projected soil organic carbon loss in response to climate warming and soil water content in a loess watershed. *Carbon balance and management, 16*(1), 1-14.

Zhao, Z., Yang, Q., Sun, D., Ding, X., & Meng, F.-R. (2020). Extended model prediction of high-resolution soil organic matter over a large area using limited number of field samples. *Computers and Electronics in Agriculture, 169*, 105172.

Zhou, G., Guan, L., Wei, X., Tang, X., Liu, S., Liu, J., Zhang, D., & Yan, J. (2008). Factors influencing leaf litter decomposition: an intersite decomposition experiment across China. *Plant soil, 311*(1-2), 61.

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., & Lausch, A. (2020). High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. *Science of The Total Environment, 729*, 138244.

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., Bumberger, J., Haase, D., & Lausch, A. (2021). Prediction of soil organic carbon and the C: N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. *Science of The Total Environment, 755*, 142661.

Zhu, M., He, Y., & He, Q. (2019). A review of researches on deep learning in remote sensing application. *International Journal of Geosciences, 10*(1), 1-11.

Zhu, W., Pan, Y., Yang, X., & Song, G. (2007). Comprehensive analysis of the impact of climatic changes on Chinese terrestrial net primary productivity. *Chinese Science Bulletin, 52*(23), 3253-3260.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8-36.

Žížala, D., Zádorová, T., & Kapička, J. (2017). Assessment of soil degradation by erosion based on analysis of soil properties using aerial hyperspectral images and ancillary data, Czech Republic. *Remote Sensing*, *9*(1), 28.