Development of NWU soil database as a precursor for a national soil database

M Sehlapelo

orcid.org/0000-0002-6101-4321

Dissertation accepted in fulfilment of the requirements for the degree *Master of Science in Environmental Sciences* at the North-West University

Supervisor: Dr PD Ramphisa

Co-supervisor: Dr GM van Zijl

Graduation October 2022

25967258

ACKNOWLEGEMENTS

I would like to thank God for giving me the wisdom, strength, and resilience to complete my thesis.

I am also thankful to my supervisors Drs. D. Ramphisa and G.M. van Zijl for the opportunity and the consistent help and guidance throughout my Masters Study.

I thank my mother Mrs. S.M. Sehlapelo and my family for the support they've shown.

I would like to thank the National Research Foundation and the Water Research Commission for funding this study. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Numbers: 121302) and the Water Research Commission Project number: C2020/2021-00455.

ABSTRACT

The availability of a high-quality database that captures the spatial variability of soil properties in South Africa, will promote the long-term sustainable use of soil in the country. In agriculture, data on the variability observed within a field is essential to meet the basic objectives of site-specific management of inputs, which will in turn lead to the increased profitability of crop production while protecting the environment and improving the quality of soil. In South Africa, the amount of soil information from Agricultural support organizations, commercial forestry and mining companies, government departments and academic institutions is broad, but not available outside of the holders of the data. The development of a soil database containing standardised soil information will not only contribute to the estimations of present and future soil productivity potential, in South Africa, but will also aid environmentalists in determining land and water limitations when conducting soil degradation risk assessments. The World Soil Information Services (WoSIS) and the Agricultural Research Council of Soil, Climate and Water (ARC-ISCW) point databases were chosen for SWOT analysis to evaluate the Strengths, Weaknesses, Opportunities and Threats that can emerge from the development of a database, and from this the NWU soil database structure was proposed. Soil point data was collected from various sources in South Africa and recorded in the NWU soil database with the accompanying soil morphological and analytical properties. Quality control measures were performed using statistical analysis including basic quality control and outlier detection to evaluate the quality of the recorded data. As a result of the evaluation of the SWOT analysis, the NWU soil database was developed in a way that it can be used to record soil data collected from various sources, recorded in different formats and created for different purposes. The final product of the NWU soil database was composed of a total of 25 sources resulting in a total of 539 soil profiles and 1518 soil horizons. The soil data was quality controlled and used to characterise the soil based on the various soil properties recorded in the soil database. Soil data from various sources can be collected, quality-controlled and recorded in a common soil database, that is complete, comprehensible, and user-friendly. This will lead to the gradual reduction of the paucity of soil data availability in South Africa.

Keywords: Database development; Structural design; Soil data; Sustainable Agriculture; Data Availability.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1 INTRODUCTION

1.1 Background of study

Soil is a multifunctional medium for storing water, filtering toxic chemicals, and transporting nutrients essential for plant growth (Pozza & Field, 2020:1). The soil's water and nutrient storage potential allows it to maintain crop yield during environmental stresses by mitigating the effect of climate change (Folberth *et al*., 2016:2). However, the intensive use of soil for agricultural production can result in the loss of soil quality leading to the degradation thereof. This is where a soil database can be of good use, it is important to evaluate the capability and condition of soil to assess the soil's optimal biophysical state to effectively carry out its functions. This evaluation involves analysis of the soil properties (Pozza & Field, 2020:5). The availability of information on the morphological and analytical properties of the soil for the implementation of environmental models for the purpose of managing environmental practices. This will increase agricultural production and mitigate soil degradation and the effects of climate change (Romero *et al*., 2012:163).

The availability of a high-quality resource dataset including soil properties from a soil profile database will also promote the use of modelling and mapping for long-term agricultural sustainability. The use of models and digital soil maps can lead to the identification of areas with good quality soil conditions for agricultural use (Batjes, 1997:9). The incorporation of advanced technology allows for experimental precision agricultural systems to represent the spatial and temporal characteristics of yield, field, crop, and soil variability within a field. These characteristics are managed through two systems: the map-based system and the sensor-based system (Zhang *et al*., 2002: 114-115). Data on the variability observed within a field is essential to meet the basic objectives of site-specific management of agricultural input, which will play a role in increasing the profitability of crop production, protect the environment and improve the quality of soil (Adamchuck *et al*., 2004:71).

Soil data obtained by different organisations is available but not outside of the holders of the data, which includes Government institutions, such as the national and provincial departments of agriculture and the Agricultural Research Centre's Institute for Soil Climate and Water (ARC-ISCW), co-operatives, forestry companies and soil scientist consultants. An exploration study carried out by ARC-ISCW revealed that more provincial soil information is placed or stored in archives (Paterson *et al.*, 2015:4). In addition, this exploration study pointed out that untapped data, not freely accessible, is not integrated into the extensive soil information system. This includes data that can be accessed from commercial and semi-commercial sources as well as public academic sources. Although there is a large amount of soil data, this readily available data are neither in the formats nor indexes required for application in Precision Agriculture (PA) and for the development of crop and environmental models (Romero *et al*., 2012:163).

The shortage of standardised soil data has not only hindered increased adoption of PA by farmers, but it also hindered the evaluation of land degradation, environmental impact analyses and development strategies for sustainable soil management (Nachtergaele *et al.*, 2010:34). A soil database can be defined as a structured set of data recorded in a computer comprising of a collection of soil samples defined by physical, chemical, and biological parameters characterising site-specific information that can enable farmers, scientists, and other stakeholders to make informed decisions about management practices on soil use (Rivera *et al*, 2015:121). The development of a soil database containing standardised soil information will not only contribute to the estimations of present and future soil productivity potential in South Africa but will also aid environmentalists in determining land and water limitations when conducting soil degradation risk assessments (Hengl *et al*., 2015:2).

In addition, the soil database will provide scientifically reliable information for planning sustainable agricultural production expansion and policy guidance to tackle emerging issues in the country. This is essential for good management of natural resources and progress towards achieving food security and sustainable agricultural production (Gebbers & Adamchuck, 2010: 828). In order to establish such a database, soil data will be collected from different entities and stored in a common Excel spreadsheet, which will include the soil properties and metadata. Quality control measures will be carried out to ensure the accuracy of the data. This data will further be used for the characterisation of soils in South Africa to identify the spatial distribution with different limitations, potentials, management needs, suitability etc. The soil database will be freely accessible to anyone wishing to use and/or record more soil data.

1.2 Problem Statement

There is a need to develop a database which captures the available spatially variable soil properties in a format that will enable land users in South Africa to access quality-assessed soil data to manage soils sustainably and productively.

1.3 Hypothesis

Soil attributes from various datasets, measured using different methods and developed for various objectives can be subjected to standardization and quality control assessments, resulting in their population into a common soil database. This will result in an accurate, comprehensible, and user-friendly soil database.

1.4 Research aims and objectives

The aim of this project is to create an NWU soil point database which is standardized, quality controlled and easily manageable and accessible.

In order to achieve this aim, the follwing objectives must be met:

1. Determining the optimal structure for the NWU soil point databse, by examining the structure of an International, World Soil Information Service (WoSIS), and a national, Institute of Soil, Climate and Water (ARC-ISCW), soil database to observe the strengths, weaknesses, opportunities and threats of the available soil databases;
2. Collect soil point data from a large variety of sources and geographic locations and populate the database structure created under objective 1;
3. Develop and perform required quality control measures to ensure the intergrity of the NWU soil point database;
4. Supply guidelines by which data could be added to the database in future;
5. Characterise the soils from the developed database using the measured soil attributes.

1.5 References

Adamchuk, V.I., Hummel, J.W., Morgan, M.T. & Upadhyaya, S.K. 2004. On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, 44: 71-91.

Batjes, N.H. 1997. A world dataset of derived soil properties by FAO-UNESCO soil unit for global modelling. *Soil Use and Management*, 13: 9-16.

Folberth, C., Skalskỳ, R., Moltchanova, E., Balkovič, J., Azevedo., L.B., Obersteiner, M. & van der Velde, M. 2016. Uncertainty in soil data outweigh climate impact signals in global crop yield. *Natural Communication*, 7, 11872, https://doi.org/10.1038/ncomms11872

Gebbers, R. & Adamchuk, V.I. 2010. Precision Agriculture and Food Security. *Science*, 327: 828-830.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamere, L. & Tondon, J.E. 2015. Mapping Soil Properties of Africa at 250m Resolution: Random Forests Significantly Improve Current Predictions. *PLoS ONE*, 10(6): 1-26.

Nachtergaele, F., Van Velthuizen, H., Verelst, L., Batjes, N., Dijkshorn, K., Van Engelen, V., Fische, G., Jones, A., Montanarella, L., Petri, M., Prieller, S., Shi, X., Teixeira, E. & Wiber, D. 2010. The harmonized world soil data. *World congress of Soil Science, Soil Solutions for a changing world*, 34-37.

Paterson, G., Turner, D., Wiese, L., Van Zijl, G., Clarke, C. & Van Tol, J. 2015. Spatial soil Information in South Africa: Situational analysis limitations and challenges. *South African Journal of Science*., 111(5/6): http://dx.doi.org/10.17159/sajs.2015/20140178

Pozza, L.E. & Field, D.J. 2020. The science of Soil Security and Food Security. Soil Security, 1. 100002, https://doi.org/10.1016/j.soisec.2020.100002

Rivera, D., Sandoval, M. & Godoy, A. 2015. Exploring soil database: a self-organizing map approach. *Soil Use Management*, 31(1), 121-131.

Romero, C.C., Hoogenboo, G., Baigorria, G.A., Koo, J., Gijsma, A.J. & Wood, S. 2012. Reanalysis of a global soil database for crop and environmental modeling. *Environmental Modelling & Software*, 35:163-170.

Zhang, N., Wang, M. & Wang, N. 2002. Precision agriculture-a worldwide view. Computers and Electronics in Agriculture, 36 (1-3):113-132, https://doi.org/10.1016/S0168-1699(02)00096-0

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

Food security can be achieved if agricultural production increases proportionally to population, through increased production or expansion of agricultural land. As there is a decline in land available for agricultural use due to land degradation and climate change, food security relies on increased production (Nhamo *et al*., 2019:1). The increase in population has a large impact on food security and it is predicted that the population will continue increasing rapidly in the coming years (Godfray & Garnett, 2014:1). Food security is largely dependent on the ability of farmers to perform precise on-site crop management to maximize farm production efficiency and profitability while protecting the environment (Adamchuk, 2004: 71). The success of increasing production depends on intensive and yet safe cultivation (including irrigation) and sustainability of agricultural land. Precision agriculture (PA) is the most effective procedure as it relies on monitoring of the spatial variation of soil to enable site specific crop management (Gebbers & Adamchuk, 2010: 828). However, PA is largely dependent on the development of comprehensive databases that captures spatial variability in soil properties (Romero *et al*., 2012:163).

Soil and land resource conservation evaluations have three main goals: to evaluate what resources are available, what land uses are feasible, and the proper management of soil. A vast number of people all over the world profitably use existing soil data to increase food production, preserve the environment and understand soil alteration over time due to climate change (Arrouys, 2014:95-96). Furthermore, up-to-date information on soil properties plays a vital role in the estimates involving crop yield, water, and nutrient dynamics, and dealing with risks in decision-making (Hengl *et al*., 2015:1). Therefore, mapping and modelling systems are determined by the availability of soil databases representing the morphological and analytical properties of soil, which are required on a scale that represents environmental impacts related to the objective of modelling systems. However, databases attributed to present data consist of limited profile data with a coarse spatial resolution (Batjes, 1997:9). Fine resolution soil maps and accurate databases are significant for adequate soil use and land management processes. Many partly completed projects which required compilation of soil databases such as the SOTER (Dijkshoorn, 2003:1) and European soil database (Panagos *et al*., 2011:434), were done to enable regional policy and decision making and revealed a variety of potential issues with integrating different datasets into a common database (Dobos *et al*., 2002:2).

2.2 Availability of soil data

2.2.1 Available soil databases

The World Soil Information Services (WoSIS) is a soil database aimed at providing georeferenced, quality-assessed soil data represented in point, polygon, and grid formats. This soil information is held at ISRIC-World Soil Information, the World Data Centre (WDC) for soils of the ICSU World Data System. The data is stored in zip file (32 MB) and consists of a manual describing detailed features of the database which is made available at doi:10.17027/isric-wdc-soils.2016003 (Ribeiro *et al*., 2020:5). In July 2016, WoSIS database recorded a total of 118,400 soil profiles, of which 17,153 were derived from Africa and 649 profiles from South Africa. This soil data was derived from the then recent soil spectral libraries and compiled soil legacy data coordinated by ISRIC such as WISE, SOTER, and African Soil profile database (Batjes, 2017:1). The African Soil Profile Database (Leenaars, 2013) consists of georeferenced, standardised, and updated legacy soil profile data obtained from Sub-Saharan Africa. The first version (1.0) of the African Soil Profile Database was developed from some 2,770 georeferenced profiles south of the Sahara derived from profile dataset of ISRIC-WISE 3. Furthermore, some data was derived from analogue reports, books, and publications available at the ISRIC World Library at international organisations, partnering countries and on the internet (Leenaars, 2013:13). The digital soil dataset from Africa Soil Profile Database is held by ISRIC and is available at: www.isric.org/data/data-download.

ISRIC-WISE database was developed from a variety of sources and harmonized according to the original (1974) and revised (1985) legend of the FAO-Unesco Soil Map of the World (Batjes, 2009:124). The latest version of the ISRIC-WISE database (3.1) is referred to as WISE 3 and consists of 10,253 soil profiles with 4,380 profiles derived from the previous versions of the database. The profiles were collected between 1925 and 2005. The data was collected from some 260 sources from 149 countries and most of the soil profiles (41%), were derived from Africa. This soil data can be obtained online at http://www.isric.org (Batjes, 2009:124).

The ARC-Institute for Soil, Climate and Water (ISCW) developed the soil science programme focused on the formation, classification and mapping of analytical soil properties linked to the sustainable management and utilisation of soils (ARC-ISCW, 2021). ARC-ISCW provides 2,595 soil profiles via the online South Africa soil dataset (AGIS). Although this data includes geographical co-ordinates, the co-ordinates were omitted and made available just for the GIS profiles shared for SOTER purposes (Leenaars, 2013:15).

In 2001 FAO and ISRIC signed an agreement to develop a Soil and Terrain (SOTER) database for Southern Africa (SOTERAF), which involved the compilation and harmonisation of the SOTER database from Botswana, Tanzania, Namibia, Mozambique, Zimbabwe, Swaziland, and South Africa. The SOTER database for the Southern African regions was compiled for the Food and Agricultural Organisation (FAO) and International Soil Reference and Information Centre (ISRIC) to improve the baseline information on natural resources (Dijkhoon, 2003:2). In July 2002, the ISCW and National Department of Agriculture in Pretoria, developed the SOTER database for South Africa (SASOTER). The SASOTER was created using the same data and format as derived from a Land Type database of ISCW, which was created from maps of the land type series and resulted in some 328 soil profiles. The final version of the database was received in November 2002, because there were errors that needed to be rectified from the first version that was created in July 2002. The soil information may be obtained from ARC-ISCW and Department of Agriculture, Pretoria (Dijkhoon, 2003:2).

2.2.2 Acquisition of soil data in South Africa

Global to local soil knowledge has often been the only missing layer of physical and biological information that lacks accuracy for predicting potential constraints for production of food (Nachtergaele *et al.*, 2010:34). Knowledge on the variation of soil properties across a wider scale plays a vital role in the impacts posed to agricultural and environmental processes (Cook *et al*., 2008: 34). The acquisition of soil data is essential within a variety of fields including health, food security, climate, and hydrological systems. Due to the variation of soil properties, the acquisition of data becomes cumbersome requiring substantial financial investments, skill, and expertise. Additionally, it becomes time consuming and labour extensive to accurately record, measure, and map these variations (Paterson *et al.*, 2015:1).

Recording of soil information acquired from existing knowledge began in the late 1930s and was published as the "Soil group and sub-groups of South Africa" (van der Merwe, 1940), resulting in the production of a national soil pattern map at a scale of 1:2500 000 (Paterson *et al*., 2015:1). Following this was the establishment of a series of soil surveys carried out to obtain soil information in some of the drier parts of the interior of South Africa for the investigation of soil erosion occurrence (Laker, 2004:346). In the 1950s, there was a need for more soil data stemming from the expansion of agricultural practices in South Africa (Phillips, 1999:254). This led to the creation of the first soil survey carried out by the sugar industry in the then Natal Province to increase the soil knowledge along the southern seaboard (Paterson *et al*., 2015:1). The second survey was carried out in Natal for the purpose of studying the wider range of soils in the Tugela Basin (Smith. 1965:106). Similarly, other surveys were carried out by the soil and Irrigation Research Institute for the Irrigation sustainability, including selected 1:50 000 map sheets (Paterson *et al*., 2015:1). Historical soil survey paper maps are valuable tools that supports soil conservation and promote sustainable land practices, especially in developing countries where digital soil data is often lacking. Many maps in these countries are kept in archives that are difficult to access (Panagos *et al*., 2011:434).

Soil information collected in the 1960s was used to create the FAO/Unesco soil map of the world and was used for over 30 years (Nachtergaele *et al.*, 2010:34). This data together with the newly updated soil data from all regions of the world encouraged the harmonisation and integration of new soil data and is available online in a Harmonised World Soil Database (HWSD) (Nachtergaele *et al.*, 2010:34). This database is currently the most detailed world soil dataset (Shangguan *et al*., 2014: 249). In the early 1970s, the national Land Type Survey took place in South Africa and lasted for over 30 years. It represented 7070 zones of terrain type, soil pattern and macroclimate zones based on the observation of about 400 000 soils (Land Type Survey Staff, 1972-2002). These observations were completed from the highly productive Highveld of the Northern Cape and the mountains of Drakensberg. Furthermore, a database containing around 2,500 modal soil profiles and 10,000 series identification were created representing soil properties across South Africa (Paterson et al., 2015:2). The Soil and Irrigation Research Institute of the Department of Agricultural Technical Services documented the distribution of soils in South Africa, this was established by creating maps and memoirs for land types at a scale of 1:250 000. This information was converted into digital format to form South Africa's soil database (Fey, 2010:5). An example of a map displaying soil information of South Africa captured in 1973 at a scale of 1: 250 0000 is shown in Figure 2.1 (Panagos *et al*., 2011). Although the Land Type survey covers the entire country, it is inadequate to use for field scale analysis (Panagos *et al*., 2011).

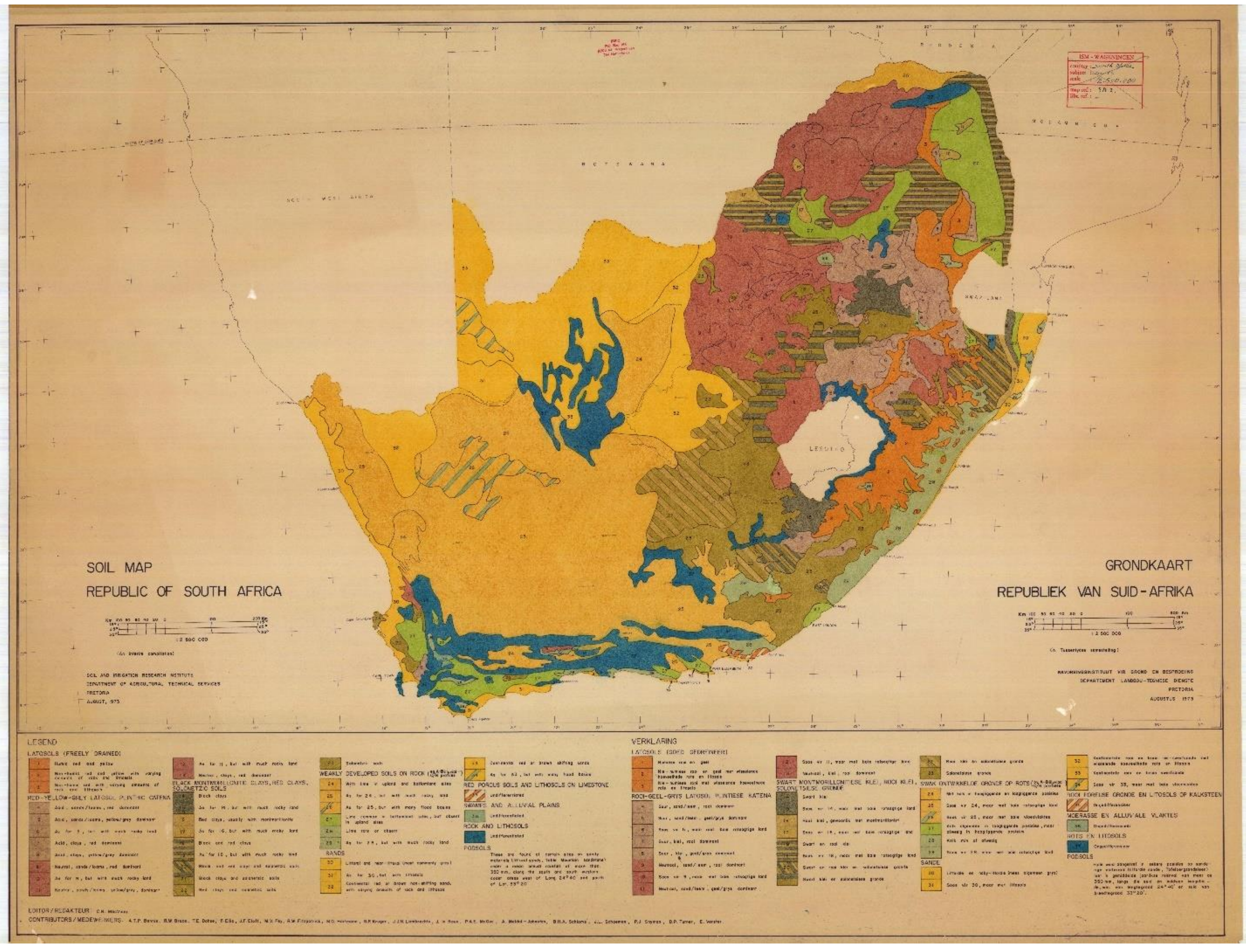


Figure 2-1: An example of a soil map displaying a collection of polygons showing areas with the same soil type (Panagos *et al*., 2011).

2.3 The importance of soil data availability

2.3.1 Soil data usage and quality

Soil is a natural resource that is non-renewable within a short space of time, it is difficult to rehabilitate and costly to restore or cultivate after erosion, physical deterioration, or chemical contamination. The deterioration and pollution of soil is often caused by pressure on land and results in a decrease in crop production capacity (Jie *et al.*, 2002: 243). The alteration of soil leading to soil degradation and loss or decrease in soil functions is influenced by human induced activities on environmental resources such as manufacturing, municipal, and agricultural activities (Olderman, 1992:19). Knowledge on the composition, characteristics, dynamics, and functions of soil as part of the landscape is the basic objective of research in soil sciences. The availability of accurate information regarding soil dynamics and characteristics obtained through analysis and description of the soil in the field is the basic prerequisite for achieving this objective (Jahn *et al.*, 2006:1). A well-designed land-use system and model development offers insight on the environment and soil management practices through the availability of soil data, such models can be used for the prevention of soil degradation and rehabilitation of deteriorated soil (Jahn *et a*l., 2006:2).

Challenges of soil data quality are not limited to the concept of uncertainty, but quality is considered a function of completeness and consistency (Veregin, 1993: 177). Furthermore, when soil measurements are perceived as good quality for a certain purpose, they may be of lesser quality for a different purpose, thus, considered as a 'fitness for use' relative concept (Hortensius & Welling, 1996:387-388). Several methods can be utilized to establish the accuracy, compatibility, and traceability of measured soil data. These include implementing and maintaining a quality management system in the laboratories and application of methods used for validation and standardization. Furthermore, the use of reference materials (that have been certified), participation and organising evaluations in inter-laboratory experiments are also essential (Theocharopous *et al*., 2004:237). Additionally, a variety of soil properties, including-positional accuracy, attribute accuracy, logical consistency, completeness, and lineage properties may be used to assess data quality (Theocharopous *et al*., 2004:238).

Fundamental to these properties, accuracy and precision are the two core essential features in data quality analysis. Reliability in data quality is increased when results are precise and accurate (Rieberro *et al*., 2015:7). The availability of a system for storing accurate soil data in such a way that the data may be accessed, integrated, and analysed in the future, for the purpose of investigating the impact on food security, the environment, and agricultural sustainability is essential. Furthermore, this system is a requirement for the formulation of policies, development planning, effective use of natural resources, and application of development programmes (Oldeman & Van Engelen, 1993:309). A comprehensive soil database compiled at a strategic and tactical level can provide soil information on the performance of different farm management strategies to enable rapid and preliminary assessments on information gained elsewhere on similar soils (Bouma, 1999:1764).

2.4 Designing a soil database

2.4.1 Types of data incorporated in soil databases

As the amount of data and data users continue to increase, a need for a systematic approach for database development is required (Fernández & Rusinkiewicz, 1993:525). The development of a database is associated with a well laid-out structure, an accurate description of its content and specification on the quality of the data to be recorded. The aim of developing a database is to ensure that the recorded data can satisfy the user's requirements, eliminate obsolete data, and perform quality control and standardization measures to provide a way of interpreting the organisation of the data. A soil database provides relevant information for national and regional planning, with a detailed risk assessment of the rate of soil degradation, land assessment and land use planning mechanisms (Oldeman & van Engelen, 1993:312). Databases used in decision support for Precision Agriculture are characterised by five main groups of soil information, primary data, secondary data, topographical data, soil fertility data and hydrological data (Figure 2.2). Primary soil data refers to data that is established through sampling and remote sensing while secondary data is data established through continuous pedotransfer functions (Bouma *et al*., 1999:1765). Soil fertility data is defined as data that represents the ability of soil to sustain growth and improve production through continuous provision of nutrients (Hartermink, 2007:1618). Hydrological data is data representing the movement, retention, and loss of water in the soil used for understanding physical and chemical processes (Vereecken *et al*., 2015:2616). Lastly, topographic data refers to digital information about the relief of a terrain (Carter, 1988: 1578).

Utilising well-documented procedures and standards to compile and process large-scale soil data is essential. The successful compilation of a systematic soil database requires reliable sources of data and standardised methods of acquiring, processing and storing this information (Batjes, 1995: 1). During a database compilation, the data is categorised into geographic and attribute data. Geographic data provides information about the location, extent and topology of each soil profile and the attribute data provides information about the soil physical and chemical properties of each soil profile. The data is processed using GIS and relational database management system (RDBMS) (Batjes *et al*., 2007:27). Once this database has been obtained it would be valid for many years and represents a once-off strategic investment (Bouma *et al*., 1999:1764-1765).

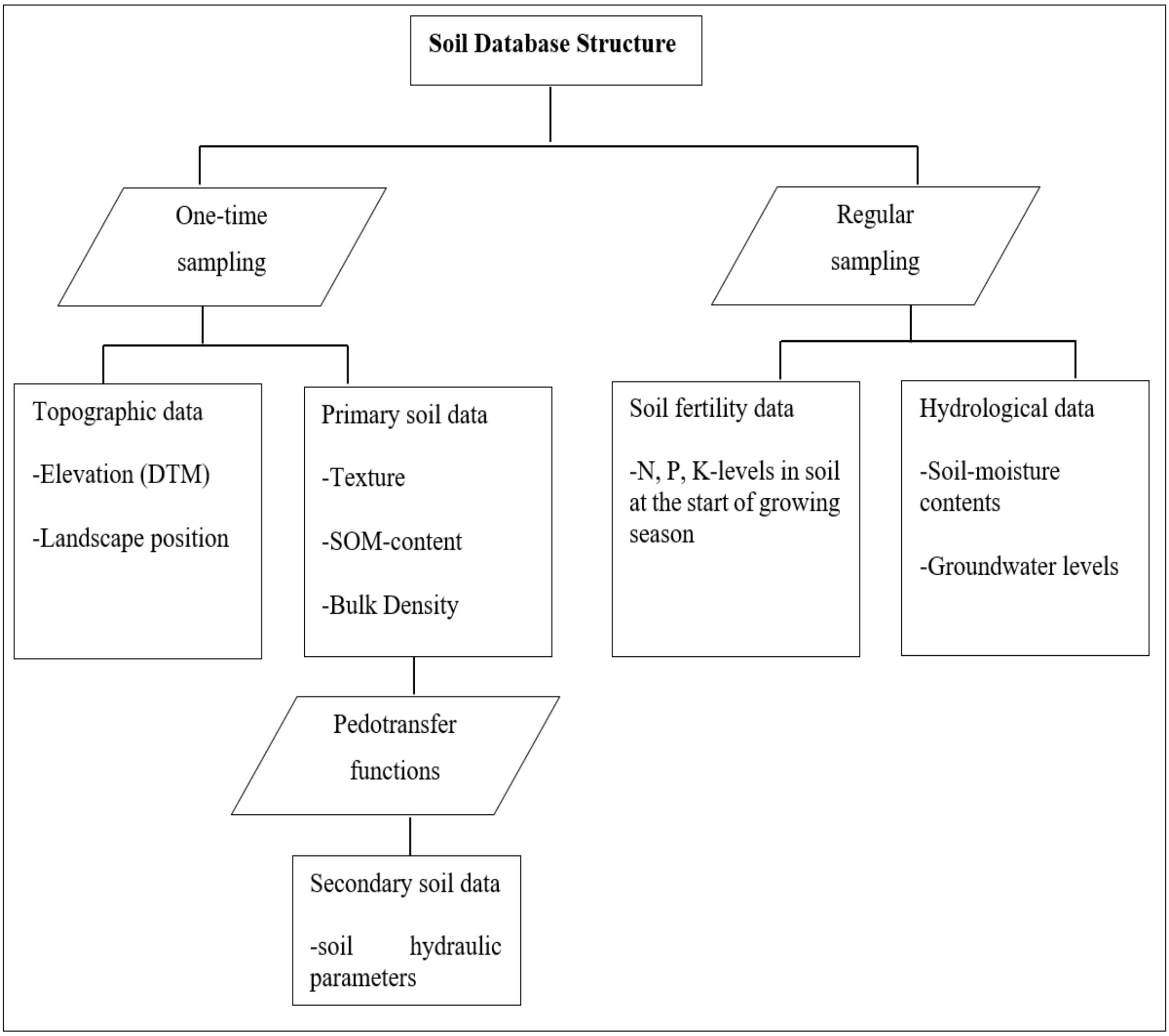


Figure 2-2: A figure displaying the five groups of soil information recorded in a soil database (Bouma *et al*., 1999:1765)

2.4.2 Procedures in database design

The process of database design involves data selection in accordance with the planned usage of the database, understanding of the expectations of the user and the development of a conceptual scheme. Furthermore, the translation of the conceptual schema into the database management systems data model (data model mapping); specifications of methods and data storage system; structure and database application are also essential in database design (Storey, 1993:25). As a result, there are several stages involved in designing a soil database. Spatial information on soil is frequently employed for decision making, while land use and soil conservation questions require increasingly accurate information on soil properties and their geographic position. As a result, the first stage in creating a soil database is deciding what the captured data will be utilized for (King *et al*., 1994:37). The second phase is gathering data from various sources, which will lead to stage 3, which involves data harmonization using data processing procedures (Shangguan *et al*., 2014:249). Stage 4 encompasses data standardization, which entails processing data from many sources to make it identical to allow data comparability during data analysis (Quevauviller, 1998:290). The resulting database should present data (Stage 5) that is easily comprehensible, accessible, and feasible (Grealish *et al*., 2004:135). The stages of data design are summarised in Figure 2.3.

Figure 2-3: The different stages involved in the conceptualisation of database design.

Understanding of the different stages involved in database design can be achieved through the analysis of different databases to compare the structure, soil information recorded, accessibility, data usage and standardisation of the data within them (Gupta, 2000:491-496). This process can be achieved through a SWOT analysis, which is a theory that refers to the strengths, weaknesses, opportunities, and threat factors (Nikolaou & Evangelinos, 2010: 26). These factors are used to identify the internal and external factors that enhance or interfere with the performance of an initiated plan or project. Strengths and weaknesses are characterised as internal factors, while the opportunities and threats are characterised as the external factors (Leigh, 2010: 115). The analyses of these factors are based on the estimation of their contributions to reaching a certain goal and the approximation of their controllability (Benzaghta, *et al*., 2021:56).

2.4.3 Limitations associated with the development of a database

When different organizations use a variety of sampling and analysis methods, this results in information that is not standardised within a database. This is because standards for different types of survey techniques and sample analysis differ, resulting in complications during the harmonisation of the database (Van Zijl & Botha, 2016:2). The development of a soil database is somewhat hampered by more limitations such as lack of laboratory analysis. Limitations associated with laboratory analysis include the significant difference between laboratory values recorded. This may be due to seasonal variations, where soil data was collected and processed at different times of the year. For example, carbon values within soils may differ due to seasonal changes. Carbon recorded during growing seasons is greater than that recorded during non-growing seasons (Van Zijl & Botha, 2016:3-5). Integrating soil data into a soil database can also be challenging when different software packages are used for capturing, managing, and storing data. In this case, semi-automatic procedures are used to transfer data into a common spreadsheet, this could lead to potential errors. Furthermore, these stages require system validation, such as verification of data position, authentication, and substantiation of data values, resulting in long processes that are costly and time consuming (Hiederer *et al*., 2006: 80). The limitations result in databases that are neither sufficiently comprehensive nor reliable to allow effective use of soil data (Lagacherie & MacBratney, 2003:4).

Incompatibility in structure and programming language exists within databases due to different interpretations of the same datasets by different people, these differences occur due to varying skills and the quality of programming languages for data modelling (Reddy *et al*., 1994: 920). Sharing and exchanging information between various database structures requires the availability of a common programming language (SQL dialect), while inconsistent semantics may result in disputes due to lack of standardization of terminology and structure definitions contained in databases (Aparicio *et al*., 2005:12). Therefore, the integration of a conceptual scheme from different databases requires identifications and comparisons of local schemes used for each database; evaluation of accuracy of the schemes; restructuring of schemes as established to meet the needs for the integration process and finally the integration of the schemes by means of standard concepts (Devogel *et al*., 2010:338)

2.5 Principles of soil data management

Soil data management is a concept that enables a computer to perform processes that can retrieve, transform, process, and classify soil data (Zhang *et al*., 2017: 5). Data management also involves the process of quality control as it leads to decreased potential of data inaccuracy (Jean-Claude *et al*., 2005). Data management and processing is essential because entities utilise several methods for collecting and analysing soil samples, as well as various ways of expressing the results. Therefore, importing all the data into the same reference system requires a significant amount of processing on both the spatial and attribute data. This procedure is referred to as "harmonisation" (Dobos *et al*., 2002:12). The computer's way to allow processes of data management is critical to the development of a good soil database (Wang *et al*., 1990:261). This is because the management of spatial data has opened new opportunities for predicting soil properties and the accompanying processes (Grunwald, 2009:196). This requires an information system that stores soil data that can be assessed, combined, modified, and accessed easily (Olderman & Van Engelen, 1993:310).

2.5.1 Statistical quality control

The extraction of knowledge and meaningful information from increasingly, diverse, and complicated soil dataset is an evident challenge for soil science (Michéli *et al*., 2016:340). Data quality control techniques can be used to build quantitative functional relationships that lead to new ideas and reveal errors in existing concepts (Cline, 1977:253). Most of the recorded data in soil surveys conducted in developing countries is categorical and statistical data (ordinal data) and this determines the type of statistical techniques that can be used on the data (Bregt *et al*., 1992:525). Processing and interpolation of soil properties using a variety of statistical techniques results in creation of soil information about a specific geographical space. Thus, soil data can be organised in Microsoft Excel software and analysed with standard statistical techniques including multiple regressions, fitted decision trees and median values analysis (Gray *et al*., 2009:310).

Outlier detection is a statistical technique used to detect errors or mistakes, in environmental sciences. This technique can be used to detect anomalies in data due to natural enrichment or anthropogenic activities affecting soil properties (Lalor & Zhang, 2001:100). Outliers can be grouped as range outliers, spatial outliers, and relationship outliers. Range outliers are detected based on the comparison between the population or majority of the sample values, where the outlier is too high or too low. Observations that are excessive in comparison to their surrounding values are known as spatial outliers and relationship outliers are observations that deviate from the predicted correlation values within a dataset (Lalor & Zhang, 2001:100). When working with large datasets, performing outlier detection is essential for maintaining data quality (Alameddine *et al*., 2010:1299). To gain a better understanding of the complicated relationships between soil qualities and environmental influences, it is necessary to characterize the geographical variability of soil attributes (Goovaerts, 1998:315).

Outlier detection can be performed in IBM SPSS (Statistical Package for the Social Sciences) software. SPSS contains powerful facilities for accomplishing and automating data management tasks such as acquiring and defining data, merging data from different sources, cleaning, aggregating, transforming, restructuring, and exporting data for analysis (Levesque, 2004:1-2). In IBM SPSS outliers can be computed using the boxplot analysis. Boxplots are used to compute the univariate data analysis and the variables are measured using ratio scales or intervals, resulting in the calculation of the mean and variation statistics (Constantin, 2017:21-23). The output of the boxplot analysis is a box bordered on the bottom by the first quartile (Q1), on the top side by the third quartile (Q3) and the median value (Q2) is delineated by a line drawn inside the box. Points plotted outside the box are regarded as potential outliers (Frigge *et al*., 1989:51).

2.5.2 Soil data characterisation

Soil databases with comprehensive morphological and analytical soil data are regarded as reliable sources of data that can enhance the possible characterisation of soils to assess soil quality and productivity in a country (Dobos *et al*., 2000:376). Soil characterisation is a concept for identifying, geographical areas of distinctive characteristics, classifying or mapping them and describing their character (Mücher *et al*., 2003:85). Soil management is often implemented due to the potential characterisation of spatial variability in soil properties, this leads to the improvement in production (Cambouris *et al*., 2006:383). Soil management using soil property data containing inherent and dynamic soil properties is essential for soil quality assessments (Oliver *et al*., 2013:130). Clay content and soil texture are classified as inherent soil properties as they are hardly altered during management practices and are therefore essential for initial soil characterisation. Soil pH, and carbon content are classified as dynamic soil properties as they are susceptible to change and are therefore considered as good indicators of change during management practices (Oliver *et al*., 2013:130).

The success of sustainable agriculture implementation is dependent on the knowledge on soils. Soil characterization provides knowledge on the surface morphology of soil that is influenced and controlled by critical environmental issues such as high surface runoff, soil degradation, floods, pollutants, and unbalanced cycles. This knowledge can be used to implement sustainable agriculture practices (Kayode *et al*., 2019:1). Soil property characterization can also be beneficial in other disciplines including geotechnical engineering, where the characterisation of properties including clay content and mineralogy are used to identify expansive soils with the potential to affect structural stability (Asuri & Keshavamuthy, 2016:29).

Simultaneous characterisation of different soil properties can be achieved by integrating digital maps, pedotransfer functions, remote sensing, spectral analysis and soil inference systems, these techniques are cost effective as compared to other methods used for soil assessments (Palm *et al.*, 2007:99). Furthermore, geostatistical techniques can also be used for soil characterisation through analysis of spatial variation of soil and crop properties across agricultural fields. This leads to effective implementation of site-specific management systems (Tola *et al.*, 2016:2). Furthermore, large datasets can be interpolated by simple geographical methods including plotting histograms for analysing distribution, data comparison and identification of extreme values, resulting in characterisation of soil data (Zhang, 2005: 1857).

2.6 Conclusion

Different organisations sampled, recorded, and stored soil data for different purposes, however, the data was not made available for public use. A variety of soil databases have been developed over the years, to capture soil information all over the world. However, the soil information for South African soils captured in these databases was not adequate to be used for field scale analyses. The availability of good quality soil data is important for the successful implementation of environmental models which enable rapid and preliminary assessment of the spatial variability of soil resulting from environmental and anthropogenic factors. As these factors contribute to the disturbance of soil to carry out its function as a medium for crop growth and sustainable agricultural production. The availability of good quality soil data is dependent on the development of a well-structured, easily comprehensible, and quality-controlled database displaying morphological and analytical soil properties. Therefore, there is a need for the development of a soil database comprising accurate soil information that is readily available for use by farmers and researchers in South Africa. However, there are limitations that can be encountered during the development of a soil database. The integration of soil data into a common database is difficult due to inconsistent semantics and lack of standardization of terminology and structure. To ensure the successful integration of soil data from different sources, there are certain procedures that need to be considered including the methods used for data collection, harmonisation, standardization, and presentation. Furthermore, the quality of a good soil database is dependent on data management and processing. This can be achieved through the implementation of basic quality control, outlier detections and correlation tests. As a result, quality-controlled data can be used for soil characterisation leading to the knowledge of soil variability in a field and successful soil management practices where necessary for sustainable agricultural production.

2.7 References

Adamchuk, V.I., Hummel, J.W., Morgan, M.T. & Upadhyay, S.K. 2004. On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, 44: 71-91.

Alameddine, I., ASCE, S.M., Kenney, M.A., Gosnell, R.J. & Reckhow, K.H. 2010. Robust Multivariate Outlier Detection Methods for Environmental Data. *Journal of Environmental Engineering*, 136: 1299-1304.

Aparicio, A.S., Farias, O.L.M. & Dos Santos, N. 2005. Applying Ontologies in the Integration of Heterogeneous Relational Databases. 11-16.

ARC-ISCW (Agricultural Research Council-).2021. Programmes of Agricultural Research Council Soil, Climate and Water. https://www.arc.agric.za/arc-iscw/Documents/ARC-ISCW%20Research%20Programmes.pdf Date of access: 01 November 2021.

Asuri, S. & Keshavamurphy, P. 2016. Expansive Soil Characterisation: an Appraisal. *INAE Letters*, 1:29-33. https://doi.org/10.1007/s41403-016-0001-9

Batjes, N.H. 1995. A homogenized soil data file for global environmental research: a subset of FAO, ISRIC and NRCS profiles (Version 1.0). Working Paper and Preprint 95/10b, International Soil Reference and Information Centre. 1-45.

Batjes, N.H. 1997. A world dataset of derived soil properties by FAO-UNESCO soil unit for global modelling. *Soil Use and Management*, 13: 9-16.

Batjes, N.H. 2009. Harmonized soil profile data for applications at continental scales: updates to the WISE database. Soil Use and Management, 25(2): 124-127.

Batjes, N.H., Al-Adamat, R., Bhattacharyya, T., Bernoux, M., Cerri, C.E.P., Gicheru, P., Kamoni, P., Milne, E., Pal, D.K. & Rawajfih, Z. 2007. Preparation of consistent soil data sets for modelling purposes: Secondary SOTER data for four case study areas. *Agriculture, Ecosystems and Environment*, 122: 26-34.

Batjes, N.H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl., T. & de Jesus, J. 2017. WoSIS: providing standardised soil profile data for the world. *Earth Systems Science Data*, 9: 1-14. doi:10.5194/essd-9-1-2017

Benzaghta, M.A., Elwalda, A., Mousa, M.M., Erkan, I. & Rahman, M. 2021. SWOT analysis applications: An integrative literature review. *Journal of Global Business Insights*, 6(1):55-73. https://www.doi.org/10.5038/2640-6489.6.1.1148

Bouma, J., Stoorvogel, J., van Alphen, B.J. & Booltink, H.W.G. 1999. Pedology, Precision Agriculture and the changing paradigm of Agricultural Research. *Soil Science Society of America Journal*, 63:1763-1768.

Bregt., A.K., Stoorvogel., J.J. & Stein, A. 1992. Mapping ordinal data in soil survey: A Costa Rican Example. *Soil Science Society of America Journal*, 56:525-531.

Cambouris, A.N., Nolin, M.C., Zebarth, B.J. & Laverdière, M.R. 2006. Soil Management Zones Delineated by Electrical Conductivity to Characterize Spatial and Temporal Variations in Potato Yield and in Soil Properties. *American Journal of Potato Research*, 83: 381-395.

Carter, J.R. 1988. Digital Representations of Topographic Surfaces. *Photogrammetric Engineering and Remote Sensing*, 54(11): 1577-1580.

Cline, M.G. 1977. Historical highlight in soil genesis, morphology and classification. *Soil Science Society of America Journal*, 41: 250-254.

Constantin, C. 2017. Using the Boxplot analysis in marketing research. *Bulletin of the Transilvania University of Brasov*. *Series V: Economic Sciences*, 10(2); 21-28.

Cook, S.E., Jarvis, A. & Gonzalez, J. P. 2008. A New Global Demand for Digital Soil Information. *Digital Soil Mapping with Limited Data*, 31-41.

Devogele, T., Parent, C. & Spaccapietra, S. 2010. On spatial database integration. *International Journal of Geographic Information Science*, 12(4):335-352.

Dijkshoorn, J.A. 2003. SOTER database for Southren Africa (SOTERSAF): technical report. IRISC-World Soil Information. http://www.isric.org/isric/webdocs/docs//SOTERSAF_Techical%20Report.pdf

Dobos, E., Bliss, N., Worstelle, B., Montanerella, L., Johannsen, C. & Micheli, E. 2002. The use of DEM and satellite data for regional scale soil database. *Agrokemia es Talajtan Hungary*. 1-12.

Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L. & Helt, T. 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. *Geoderma*, 97: 367-39.

Fernández, R.N. & Rusinkiewicz, M. 1993. A conceptual design of a soil database for a geographical information system. I*nternational Journal of Geographic Information Systems*, 7(6): 525-539.

Fey, M.V. 2010. A short guide to the soils of South Africa, their distribution and correlation with World Reference Base soil groups. 2010 19th World Congress of Soil Science, Soil Solutions for a changing World, 32-35.

Frigge, M., Hoaglin, D.C. & Iglewicz, B. 1989. Statistical Computing: Some Implementations of the Boxplot. *The American Statistician*, 43 (1): 50-54.

Gebbers, R. & Adamchuk, V.I. 2010. Precision Agriculture and Food Security. *Science,* 327: 828-830.

Godfray, H.C.J. & Garnett, T. 2014. Food security and sustainable intensification. *Phil. Trans. R. Soc. B*, 369 : 1-10.

Goovaerts, P. 1998. Geostatistical tools for characterizing the spatial variability of microbiological and physio-chemical soil properties. *Biological Fertility Soils*, 27 :315-334.

Gray, J.M., Humphreys, G.S. & Deckers, J.A. 2009. Relationships in soil distribution as revealed by a global soil database. *Geoderma*, 150: 309-323.

Grealish, G., King, P., Omar, S. & Roy, W. 2004. Geographic information system and database for the soil survey for the State of Kuwait-design and outputs. *Kuwait Journal of Science and Engineering*, 31 (1);135-148.

Grunwald, S. 2009. Multi-criteria characterization of recent digital soil mapping and modelling approaches. *Geoderma*, 152: 195-207.

Gupta, R. 2000. SWOT analysis of geographic information: The case of India. *Current Science*, 79 (4): 489-498).

Hartemink, A.E. 2007. Soil Fertility Decline: Definitions and Assessment. *ISRIC-World Soil Information*, 1618-1621.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamere, L. & Tondon, J.E. 2015. Mapping Soil Properties of Africa at 250m Resolution: Random Forests Significantly Improve Current Predictions. *PLoS ONE*, 10(6): 1-26.

Hiederer, R., Jones, R.J.A. & Daroussin, J. 2006. Soil Profile Analytical Database for Europe (SPADE): Reconstruction and Validation of Measured Data (SPADE/M). *Geografisk Tidsskrift-Danish Journal of Geography*, 106 (1): 71-85.

Hortensius, D. & Welling, R. 2008. International standardization of soil quality measurements. Communications in Soil Sciences and Plant Analysis, 27 (3-4): 387-402. https://doi.org/10.1080/00103629609369563

Jahn, R., Blume, H.P., Asio, V.B., Spaargaren, O. & Schad, P. 2006. Guidelines for soil description. 4th edition. Rome, Italy: FAO.

Jean-Claude, G., Christophe, C., Gilles, B. & Bernard, J.2005. EcoPlant: A fore site database linking floristic data with soil and climate variables. *Journal of Vegetation Science*, 16;257-260.

Jie, C., Jing-zang, C., Man-zhi, T. & Zi-tong, G. 2002. Soil degradation: a global problem endangering sustainable development. *Journal of Geographical Sciences*, 12:243-252.

Kayode, O.T., Aizebeokhai, A.P. & Odukoya, A.M. 2019. Soil characterisation for precision agriculture using remotely sensed imagery in southeastern Nigeria, *Journal of Physics: Conference Series*, 1-8.

King, D., Daroussin, J. & Tavernier, R. 1994. Development of a soil geographic database from the soil map of the European Communities. *Catena*, 21: 37-56.

Lagacherie, P. & McBratney, A.B. 2003.Spatial soil information systems and spatial inference systems: Perspectives for Digital Soil Mapping. *Developments in Soil Science*, 31:3-22.

Laker, M.C. 2004. Advances in soil erosion, soil conservation, land suitability evaluation and land use planning research in South Africa, 1978-2003. *South African Journal of Plant and Soil*, 21(5): 345-368.

Land Type Survey Staff. 1972-2002. 1:250 000 scale Land Type Survey of South Africa. Pretoria: ARC Institute for Soil, Climate and Water.

Lalor, G.C. & Zhang, C. 2001. Multivariate outlier detection and remediation in geochemical databases. *The Science of the Total Environment*, 281:99-109.

Leenars, J.G.B. 2013. African Soil Profiles Database Version 1.1. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). ISRIC Report 2013/03. 1-160.

Levesque, R. 2004. SPSS Programming and Data Management. 2nd edition. Chicago: SPSS.

Michéli, E., Láng, V., Owens, P.R., McBratney, A. & Hempel, J. 2016. Testing the pedometric evaluation of taxonomic units on soil taxonomy- A step in advancing towards a universal soil classification system. *Geoderma*, 264: 340-349.

Mücher, C.A., Bunce, R.G.H., Jongman, R.H.G., Klijn, J.A., Koomen, A.J.M., Metzger, M.J. & Wascher, D.M. Identification and characterisation of environments and landscapes in Europe. Wageningen: Alterra (Altera-rapport 832)-119.

Nachtergaele, F., Van Velthuizen, H., Verelst, L., Batjes, N., Dijkshorn, K., Van Engelen, V., Fische, G., Jones, A., Montanarella, L., Petri, M., Prieller, S., Shi, X., Teixeira, E. & Wiber, D. 2010. The harmonized world soil data. *World congress of Soil Science, Soil Solutions for a changing world*, 34-37.

Nhamo, L., Matchaya, G., Mabhaudhi, T., Nhlengethwa, S., Nhemachena, C. & Mphandeli, S. 2019. Cereal Production Trend under Climate Change: Impacts and Adaptation Strategies in Southern Africa. *Agriculture*, 9 (30): 1-19.

Nikolaou, I.E., & Evangelinos, K.I. 2010. A SWOT analysis of environmental management practices in Greek Mining and Mineral Industry. *Resources Policy*, 35: 226-234.

Oldeman, L.R. 1992. Global Extent of Soil Degradation. ISRIC Bi-Annual Report. 19-36.

Oldeman, L.R. & Van Engelen, V.W.P. 1993. A world soils and terrain digital database (SOTER) - An improved assessment of land resources. *Geoderma*, 60: 309-325.

Palm, C., Sanchez, P., Ahamed, S. & Awiti, A. 2007. Soils: A Contemporary Perspective. *Annual Review of Environment and Resources*, 32: 99-129.

Paterson, G., Turner, D., Wiese, L., Van Zijl, G., Clarke, C. & Van Tol, J. 2015. Spatial soil Information in South Africa: Situational analysis limitations and challenges. *South African Journal of Science*., 111(5/6): 1-7.

Panagos, P., Jones, A., Bosco, C., Senthil Kumar P.S. 2011. European digital archive on soil maps (EuDASM): preserving important soil data for public free access. *International Journal of Digital Earth*, 4 (5), pp. 434-443., DOI:10.1080/17538947.2011.596580.

Phillips, S.T. 1999. Lessons from the Dust Bowl: Dryland Agriculture and Soil Erosion in the United States and South Africa, 1900-1990. *Environmental History*, 4(2): 245-266.

Quevauviller, P. 1998. Operationally defined extraction procedures for soil and sediment analysis I. Standardization. *Trends in analytical chemistry*, 17 (5): 289-297.

Reddy, M.P., Prasad, B.E., Reddy, P.G. & Gupta, A. 1994. A Methodology for Integration of Heterogeneous Databases. *IEEE Transactions on Knowledge and Data Engineering*, 6: 920-933.

Romero, C.C., Hoogenboo, G., Baigorria, G.A., Koo, J., Gijsma, A.J. & Wood, S. 2012. Reanalysis of a global soil database for crop and environmental modeling. *Environmental Modelling & Software*, 35:163-170.

Rossiter, D.G. 2004. Digital soil resources inventories: status and prospects. *Soil Use and Management*, 20: 296-301.

Shangguan, W., Dai, Y., Duan, Q., Liu, B. & Yuan, H. 2014. A global soil data set for earth system modelling. *J. Adv. Model. Earth Syst*., 6: 249-263.

Smith, E.T. 1965. The approach to regional planning in Natal. *Koers Journal*, 22: 2, 101-118. https://doi.org/10.4102/koers.v33i2.1353

Storey, V.C. & Goldstein, R.C. 2003. Knowledge-Based Approaches to Database Design. *MIS Quarterly*, 17(1): 25-46.

Theocharopoulos, S.P., Mitsios, I.K. & Arvanitoyannis, I. 2004. Traceability of environmental soil measurements. *TrAC Trends in Analytical Chemistry*, 23(3): 273-251. https://doi.org/10.1016/S0165-9936(04)00317-6

Tola, E., Al-Gaadi, K.A., Madugundu, R., Zeyada, A.M, Kayad, A.G. & Biradar, C.M. 2016. Characterization of spatial variability of soil physicochemical properties and its impact on Rhodes grass productivity. *Saudi Journal of Biological Sciences,* 1-9.

Van der Merwe, C.R. 1940. Soil groups and sub-groups of South Africa. Science Bulletin 231. Pretoria: Department of Agriculture and Forestry.

Van Zijl, G.M. & Botha, J.O. 2016. In pursuit of a South African national soil database: potential and pitfalls of combining different soil data sets. *South African Journal of Plant and Soil*, 1-7.

Vereecken, A., Huisman, J.A., Franseen, H.J.H., Bruggeman, N., Bogena, H.R., Kollet, S., Javaux, M., van der Kruk, J. & Vanderborght, J. 2015. Soil hydrology: Recent methodological advances, challenges, and perspective. *Water Resour. Res*., 51 :2616-2633.

Zhang, C., Manheim, F.T., Hinde, J. & Grossman, J.N. 2005. Statistical characterization of a large geochemical database and effect of sample size. *Applied Geochemistry*, 20 :1857-1874.

Zhang, X., Zhang, J., Li, L., Zhang, Y. & Yang, G. 2017. Monitoring Citrus Soil Moisture and Nutrients Using an loT Based Systsem. *Sensors*, 17(447). doi:10.3390/s17030447

CHAPTER 3 NWU SOIL DATABASE STRUCTURE

3.1 Introduction

A lack of publicly available soil data has been named as a factor preventing production sustainability and environmental protection, among other factors (Wösten *et al*., 1999:169). To tackle the paucity of available soil data, an initiative must be taken to bring together all available soil data into a unified soil database. However, the development of a soil database requires a robust database structure design that ensures that data obtained from different sources is adequately stored and readily accessible (Grealish *et al*., 2004:135). Database design is defined as the development of the database structure aimed at describing the contents and limitations to be imposed on the validity of the information. The development of an information system typically requires a systematic approach when designing a database (Fernández & Rusinkiewicz, 1993:525).

The aim of this chapter is to examine the structure of the international World Soil Information Servives (WoSIS) and a national Institute of Soil, Climate and Water (ARC-ISCW) soil databases to observe the strengths, weaknesses, opportunities, and threats thereof and propose the structural design of the NWU soil database based on these observations. It is essential to examine the various soil database structures in order to determine which structure will best fit the requirements for a robust soil database design. This is attributed to the reason that the structural design of a database is dependent on the user requirements including the results and information that can be obtained from the dataset (Melikhova *et al*., 2019:1185) . In this chapter the structural design of the NWU soil database will be discussed including the methods and materials involved in achieving the structure.

3.2 Materials and methods

To meet the aim of this chapter, two databases were obtained, namely the international and a national database which were subjected to the SWOT analysis (strengths, weakness, opportunities, and threats) to observe data legibility, quality, accessibility, usage, and standardisation for the different databases. The two databases are the WoSIS (The World Soil Information Service) database and the ARC-ISCW (Agricultural Research Council of Soil, Climate and Water) database, respectively. The databases were chosen for the analysis because each database was used to record data at different geographical spaces and for different purposes resulting in very distinct structures. This enabled a broader observation and comprehension. The Strengths, Weakness, Opportunities, and Threats of each database were evaluated and from this the structure of the NWU database was proposed.

3.2.1 Obtaining the WoSIS and ARC-ISCW databases

WoSIS database was downloaded from the ISRIC Data hub saved in a zip file under "download and links" on the site. When downloaded, the database is saved on the PC as a WoSIS_2019_September folder. In this folder there are six files; a *Readme First_WoSIS_2019dec04* pdf document with all the definitions of the codes used in the database describing the soil attributes; a *wosis_201909 geoPackage (gpkg)* file; and four tsv. files: *wosis_201909_attributes.tsv*, *wosis_201909_profile.tsv*, *wosis_201909_layers_chemicals.tsv* and *wosis_201909_layer.tsv*.

ARC-ISCW database was obtained by contacting Garry Paterson, the soil pedology research manager at the ARC-ISCW. This database consists of 38 separate PDF documents, where a single soil profile is described on two separate pdf documents. One document displayed morphological soil properties, while the other document displayed the analytical soil properties of the soil profile.

3.2.2 SWOT analysis

To develop a good soil database structure, strengths, weaknesses, opportunities, and threats of two databases were evaluated. The criteria used for the SWOT analysis included the assessment of:

(i) The structure used to store the data (legibility):

- What software is used to store the data, and in what format is the data recorded for representation - is the data recorded in well-organized tables with each attribute represented in a single column, or is it contained in text?
- The ease to navigate through the database to find soil information.

(ii) Soil data (properties) provided (data quality):

- Which soil attributes are used to provide information about the soil?
- Are soil attributes like geographic co-ordinates, units of measurement, and methods provided in the dataset?

(iii) Accessibility of database (data accessibility):

- How and where the database can be obtained?

(iv) Data usage:

- Is the data easily retrievable from the database?
- Can the data be copied using the "copy" and "paste" option to maintain accuracy or does the data have to be manually copied into other software for use?
- How easy or difficult it is to make sense of the data and to use this data in other software packages?
- How easy or difficult is it to add more soil data into the database?

- Can the data go through quality control procedures, and can it be added into programs like GIS and used to create maps models for environmental assessments?

(v) Standardization of data

- Did the source provide information on the units and methods of measurements used to obtain the data and the soil classification system thereof?
- Are the lab methods used standard and can new methods be easily added?
- Can this data be easily converted to other units of measurements or methods of measurements and integrated in other databases?

The goal of this analysis was to determine what strengths make up a good database structure and which weaknesses should be avoided while designing this structure. Furthermore, the threats that may result in restrictions to achieving the data usage objective, as well as the opportunities that could address gaps and generate new activities for the user were also determined.

3.2.3 Creating the NWU soil database structure

The criteria set forth was used as a guideline to propose a suitable structure for the NWU soil database. Based on this criterion, the SWOT analysis was done to evaluate the strengths, weaknesses, opportunities, and threats of the WoSIS and ARC-ISCW databases. The aim was to develop the NWU soil database structure that incorporates the strengths and opportunities and negates the weaknesses and threats established during the SWOT analysis. Based on the criteria a good soil database is characterised by its legibility, data quality, accessibility of data usage and standardisation.

As stipulated in the criteria, the legibility of a database is affected by the software used for storing data, the format used for recording the data and the ease to navigate through the database to find data. Data quality of a database is affected by the soil information and metadata provided. Accessibility of a database is affected by the ease with which the data can be obtained. Data usage of a database is affected by the ease of data retrieval from the database, availability of manuals on how to use the database, the ease of understanding the content of the database and the ease of data expansion of the database and performing quality control. Standardisation of data within a database is affected by the availability of units and methods of measurements and classification system used for the data, the use of standard lab methods, the ease of converting units and methods of measurements. Therefore, the advantages and opportunities ensuring the legibility, data quality, accessibility, data usage and standardisation of a database were incorporated into creating the NWU structure and the weaknesses and threats observed were negated.

3.3 Results and discussion

3.3.1 Database Description

3.3.1.1 WoSIS

The WoSIS database is made up of seven separate files, as mentioned above. A Readme First_WoSIS_2019dec04 pdf document, a WOSISprocedureManual_2020nov17web pdf document a wosis_201909 geoPackage (gpkg) file and four tsv files: wosis_201909_attributes.tsv, wosis_201909_profile.tsv, wosis_201909_layers_chemicals.tsv and wosis_201909_layer physical.tsv. The Readme First_WoSIS_2019dec04 document contains an overview of the information each document provides, the field names in each document and the description thereof. Furthermore, the document provides guidelines on the usage of data while the WoSISprocedureManual_2020nov17web provides guidelines for querying and handling the data.

The wosis_201909_attributes.tsv file (Figure A1.1 in the Appendix) contains codes used to represent each attribute, to indicate whether the attribute is a horizon or a site property. The codes also represent units of measurements, the number of profiles or layers and a brief description of each attribute, including the implied uncertainty for each property. The wosis_201909_profile.tsv file (Figure A1.2 in the Appendix) contains the unique profile ID, the source of the data, country ISO code and name, accuracy of geographical co-ordinates, latitude, and longitude in the format of the World Geodetic System 1984 (WGS 1984), point geometry of the location of profile, samples' maximum depth of soil calculated and information on the soil classification system.

The wosis_201909_layer_chemical.tsv (Figure A1.3 in the Appendix) and wosis_201909_layer_physical.tsv files (Figure A1.4 in the Appendix) contain information on soil chemical and physical attributes respectively. The above files have tab-delimited fields with double quotation marks as text delimiters. Since the data was too vast to fit in one file, these attributes were saved in two separate files. However, all the files could be imported into a Standard Query Language (SQL) database or statistical software such as R and joined into a single file using the distinctive profile_id.

The first two columns of Table 3.1 contain information as recorded in the wosis_201909_layer_chemical.tsv and wosis_201909_layer_physical tsv. files and the terminology used to describe each attribute. The last two columns of the table contain information as recorded in the wosis_201909_attributes.tsv file and the terminology used to describe each attribute. Table 3.2 represents all the soil attributes recorded according to different soil classification systems as recorded on the wosis_201909_profiles.tsv. file. These include the World Soil Reference Base (WRB), USDA Soil Taxonomy (UST) and the Food and Agricultural Organisation (FAO) classification system. A table containing all the soil properties, units of measurements and description is provided on the ISRIC World Soil Information site www.isric.org/data/data-download.

Chemical properties that were measured in WoSIS include the calcium carbonate equivalent total (tceq), cation exchange capacity-buffered at pH7 (ecph7), cation exchange capacity-buffered at pH8 (cecph8), effective cation exchange capacity (ecec), electrical conductivity- ratio 1:2 (elco20), electrical conductivity- ratio 1:2.5 (elco25), electrical conductivity- ratio 1:5 (elco50), organic carbon (orgc) and pH $CaCl_2$ (phca). Physical properties that were measured include the bulk density fine earth-33kPa (bdfi33), bulk density fine earth- air dry (bdfiad), bulk density fine earth-field moist (bdfifm), bulk density fine earth- oven dry (bdfiod), bulk density whole soil-33 kPa (bdws33), bulk density whole soil- air dry (bdwsad), bulk density whole soil-field moist (bdwsfm), and clay %.

Table 3-1: Chemical, physical and types of attributes as recorded in WoSIS database and the descriptions of the term used.

Wosis_201909_layer_chemical.tsv and Wosis_201909_layer_physical.tsv		**Wosis_201909_attributes.tsv**	
Profile_id	Identifier for profile	Code	Four to six letter codes for each attribute
Profile_layer_id	Unique identifier for layer for given profile (primary key)	Type	Whether the attribute is a site or horizon property
Upper_depth	Upper depth of layer (horizon)	Attribute	Name of recorded attribute
Lower_depth	Lower depth of layer	Unit	Unit of measurement
Layer_name	Name of horizon	Profiles	The number of files respectively layers
Litter	Flag (Boolean), indicating whether this is considered a surficial litter layer	Layers	
xxxx_value	Array listing all measurement values for soil property 'xxxx'	Description	A brief description of each attribute
xxxx_value_avg	Average for 'xxxx_value'	Accuracy	Inferred uncertainty for each property
xxxx_method	Array listing the method descriptions for each value		
xxxx_date	Array listing the date of the observation for each value		
xxxx_dataset_id	Abbreviation for source data set		
xxxx_profile_code	Code for given profile		
xxxx_license	License for given data, as indicated by the data provider		

Table 3-2: Profiles as recorded in WoSIS database and the descriptions thereof

Wosis_201909_profiles.tsv		Wosis_201909_profiles.tsv	
Profile_id	Profile key and foreign key that refers to table "profile"	Cwrb_reference_soil_group_code	Code for WRB group (in given version of WRB)
Dataset_id	Identifier for source data set	Cwrb_reference_soil group	Full name for reference soil group
Country_id	ISO code for country name	Cwrb_prefix_qualifier	Name for prefix (e.g., for WRB1988) resp. principal qualifier (e.g., for WRB2015)
Country_name	Country name	Cwrb_suffix_qualifier	Name for prefix (e.g., for WRB1988) resp. supplementary qualifier (e.g., for WRB2015)
Geom_accuracy	Accuracy of the geographical co-ordinates in degrees.	Cstx_version	Version of USDA Soil Taxonomy (UST)
Latitude	Latitude in degrees (WGS84)	Cstx_reference_soil_group_code	Code for UST group (in given version of UST)
Longitude	Longitude in degrees (WGS84)	Cstx_reference_soil_group	Full name of reference group
Dsds	Maximum depth of soil described and sampled (calculated)	Cstx_prefix_qualifier	Name of prefix resp. supplementary qualifier
Cfao_version	Version of FAO legend	Cstx_version	Version of USDA Soil Taxonomy (UST)
Cfao_major _group code	Code of major group-depending on the version given in the Legend	Cstx_order_name	Name of UST order
Cfao_major_group	Name of major group	Cstx_suborder	Name of UST suborder
Cfao_soil_unit_code	Code for soil unit	Cstx_great_group	Name of UST great group
Cfao_soil_unit	Name of soil unit	Cstx_subgroup	Name of UST subgroup

Cwrb - (Soil Classification-World Reference Base for Soil Resources)
Cstv - (Soil Classification-USDA Soil Taxonomy)
USDA- (United State Department of Agriculture)
Cfao- (Soil Classification-Food Agriculture Organisation)

3.3.1.2 ARC-ISCW

Soil data in the ARC-ISCW is recorded and stored using MS Access, with descriptive products supplied in the form of several PDF documents, with two separate documents containing soil morphology descriptions (Figure 3.1) and soil analytical properties (Table 3.3). Each document holds information for a single profile. Soil morphology descriptions displayed in Figure 3.1 include profile no., map no., co-ordinates, land type no., climate zone, altitude, terrain unit, slope/shape, aspect and microrelief, and parent material. Furthermore, there is also information displayed on the soil form and series, surface rockiness, occurrence of floods, wind and water erosion, water table, name of observer, date of observation, weathering of underlying material, soil horizons, depth, description, and diagnostic horizons.

Soil analytical properties displayed in Table 3.3 include particle size, Citrate-bicarbonate-dithionite (CBD) extracted metals, mineralogy, Organic Carbon (OC), Cation Exchange Capacity (CEC), exchangeable cations, extractable acidity, pH, soluble cations, saturation %, Sodium Adsorption Ratio (SAR), resistivity, Electrical Conductivity (EC), micro nutrients, P-status, P-sorption, Resilient Modulus (MR), relative soil saturation (AWR), liquid limit, plastic limit, linear shrinkage and plasticity index.

ARC • LNR

SOIL PROFILE DESCRIPTION

NATIONAL SOIL PROFILE NO : 1718

Map/photo : 2726 Kroonstad

Latitude + Longitude: 27° 1' 0" / 27° 3' 24"

Land Type No : Bd13

Climate Zone 31S

Altitude : 1356 m

Terrain Unit: Footslope

Slope: 1 %

Slope Shape : Convex

Aspect : South

Microrelief : None

Parent Material Solum

Underlying Material :

Soil form and serie Glencoe weltevrede

Surface rockiness None

Surface stoniness : None

Occurence of flooding None

Wind erosion : None

Water Erosion None

Vegetation / Land use

Water table : None

Described by : R.W. Bruce

Date Described : 7704

Weathering of underlying material

Horizon	Depth (mm)	Description	Diagnostic horizon
O	0 - 270		
A1	270 - 550	Overburden: Moist state; moist colour: dark yellowish brown 10YR4/4; texture: fine sand; structure: massive; consistence: friable.	
B21	550 - 1000	Moist state; moist colour: yellowish brown 10YR5/6; texture: loamy medium sand; structure: single grain; consistence: friable; few ; few fine <2-6mm sesquioxide concretions; abrupt wavy transition.	Yellow-brown apedal
B22	1000 - 1100	Ferricrete: texture: loamy fine sand.	

Figure 3-1: The structural format of the ARC-ISCWC soil morphological description dataset.

Table 3-3: The structural format of the ARC-ISWC soil analytical properties.

ARC • LNR

INSTITUTE FOR SOIL, CLIMATE AND WATER
INSTITUUT VIR GROND, KLIMAAT EN WATER

Described by : R.W. Bruce **Date described** 7704

Nasionale Nr / National No : 1718 **Breedtegraad / Latitude:** 27° 1' 0" **Landtipe / Land Type:** Bd13

Soil form and series Glencoe weltevrede Gc14 **Kaart / Map :** 2726 Kroonstad **Lengtegraad / Longitude:** 27° 3' 24" **Klimaatsone / Climate Zone** 31S

Lab nr/no	Diepte Depth	Hor	Particle Size < 2mm							CBD			**Mineralogie / Mineralogy	
			coSand	meSand	fiSand			Silt	Clay	Fe	Al	Mn	2 - 50 µm	< 2 µm
	mm		%							%				
C 4362	270 - 550	A1	1.4	33.3	57.6			0.6	5.7	0.37	0.04	0.00	Qz5,Fs1,Tc1	Kt5,Tc5,Mi4,Qz3,Ch(2:1)2,Vm2
C 4363	550 - 1000	B21	6.0	30.9	49.0			4.8	8.8	0.54	0.07	0.00	Qz5,Fs1,Tc1	Kt5,Mi4,Tc4,Qz3,Ch(2:1)2
C 4364	1000 - 1100	B22	4.4	28.8	52.9			3.3	9.0	2.11	0.22	0.00	Qz5,Tc2,Fs1,Kt1	Kt5,Tc5,Mi4,Qz3,Ch(2:1)2

Lab nr/no	Org C	KUK CEC	Katione / Cations * Uitruilbaar / Exchangeable # Ekstraheerbaar / Extractable						Ekstr. suurheid Extractable acidity		pH		Versadigingsekstrak / Saturation Extract Oplosbare Katione / Soluble Cations								Vers. Satur.	NAV SAR	Weerst. Resist.	EG EC
				Na	K	Ca	Mg	Tot	Total	Al	H2O	KCl	Na	K	Ca	Mg	Na	K	Ca	Mg				
	%	cmol (+)/kg soil							cmol (+)/kg soil				cmol (+)/dm³				cmol (+)/ kg soil				%		ohms	mS/m
C 4362	0.30	1.70	#	0.00	0.10	0.40	0.10	0.60	0.40	0.10	5.10												15000	
C 4363	0.20	2.30	#	0.00	0.10	0.20	0.10	0.40	0.70	0.40	4.80												11000	
C 4364	0.20	3.00	#	0.00	0.00	1.90	0.10	2.00	0.20		5.80												8300	

Lab nr/no	Mikrovoedingselemente Micro nutrients					P			Waterretentiwiteit Water Retentivity				MR	AWR	Atterberg Limiete / Limits			
						Status		Sorpsie Sorption							Vloeigrens Liquid limit	Plastisiteitsgrens Plastic limit	Liniêre krimping Linear shrinkage	Plastisiteitsindeks Plasticity index
	Zn	Mn	Cu	Co	B	P	Bray		-33 kPa	-80 kPa	-500 kPa	-1500 kPa						
	mg / kg					mg/kg		%	%				kPa		%			
C 4362	0.09	0.80	0.65	0.14	0.47	0.90		27.04	3.70	3.20	2.60	2.40		0.20				
C 4363	0.44	0.10	0.69	0.03	0.32	0.40		63.04	5.50	4.70	3.80	3.60	12.10	0.10				
C 4364	0.18	0.30	0.78	0.15	0.32	0.70		87.49	8.60	6.80	5.20	4.80	11.30	0.10				

Lab nr/no	METODE GEBRUIK / METHOD USED							
	Status		Sorpsie / Sorption	Boron	KUK / CEC	Katione / Cations	CBD	Mikro Elemente Micro Elements
	P	P-Bray						
C 4362	#Error		#Error	#Error	#Error	#Error	#Error	#Error
C 4363	#Error		#Error	#Error	#Error	#Error	#Error	#Error
C 4364	#Error		#Error	#Error	#Error	#Error	#Error	#Error

3.3.1.3 WoSIS

Table 3.4 gives the SWOT analysis of the WoSIS database, while discussion of the different points will follow in the text. Based on the structural analysis of the WoSIS database and criteria set fourth, the following SWOT analysis was used to propose a structure for the NWU soil database.

3.3.1.3.1 Strengths

Legibility

Data can be imported into SQL and statistical software

All the soil information that makes up the WoSIS database is recorded in four different files, wosis_201909_attributes.tsv, wosis_201909_profile.tsv, wosis_201909_layers_chemicals.tsv and wosis_201909_layer physical.tsv. These files can be imported into SQL (Standard Query Language) database or statistical software such as R where they may be joined using the profile_id. Importation of the dataset allows for handling and querying of the data as required by the user. Guidelines of this procedure is provided in the WoSIS Procedures Manual (Ribeiro *et al*., 2020 46).

Data quality

Availability of information on database

There are articles, reports and documentations about the database, the contents include definitions of soil properties and a set of attributes that can be used to express a description on a measurement. Downloads and links, contact details for enquiries and the source of the dataset can also be found on the ISRIC World Soil Information page www.isric.org/data/data-download.

Accessibility

Database available online

The WoSIS database is readily available online. The ISRIC World Soil Information page provides information on the product- standardised dataset as derived from WoSIS database, accessibility of the database and tutorials on how to use the R software.

Standardization of data

Methods described in columns

The methods of measurement are readily described in the same table and column as the soil attribute, this results in a comprehensible and user-friendly database. There is also a specification of the soil classification used for each set of data.

Data Usage

Availability of link table

Linking the different soil property tables allows easier navigation through the database, the user does not need to navigate back and forth through each file to use the data. Furthermore, the ease with which each TSV (tap separated values) file can be converted into an excel spreadsheet improves the tabular presentation of soil attributes and the use of the database.

3.3.1.3.2 Weaknesses

Legibility

Large amount of data

There is a large volume of data recorded in many columns, which makes the data appear clustered resulting in poor data presentation and difficulty working through the database.

Table 3-4: The SWOT analysis of the WoSIS database.

S	W
• TSV file can easily be imported into an SQL database or statistical software such as R, after which they may be joined using the unique profile id. • Availability of information on the database. • Database readily available online. • Methods used are described in columns. • Availability of link table to connecting the profile to a given source.	• Large amount of data, incomprehensible and difficult to go through all the data. • No specified units of measurements in tables • The data was saved on a TSV file, unfamiliar to most potential users. • Data spread over different files. • Unfamiliar column headings.
O	**T**
• Smaller geographical based datasets could improve usage considerably	• First time users could be deterred by unfamiliar file types and headings, as well as various documents to be consulted. • Lack of quality control due to missing units of measurements.

SQL- (Standard Query Language)

Standardisation of data

No specified units of measurements

Scientific measurements generally adhere to the International System of Units (SI units). It is important to always include units when recording data, doing calculations, and reporting results so data can make sense to the people it is shared with. Knowing the units of measurement for specific soil properties while analysing results recorded for the property is more effective than looking for this information in other files which proves to be time consuming.

Data Usage

Data saved in TSV files

WoSIS dataset is saved in TSV file format which opens in notepad making it is difficult to read and edit, i.e., to get the data to display in tables with headings the files must be opened in excel. While this process may be simple to experienced users it limits the number of users because not everyone is experienced with working on different software systems and converting data from one software to the next. It is evident that the WoSIS database was set up for experienced users.

Data spread over different files.

The data is spread over four soil attribute files, which makes it difficult to simultaneously use data from the same observation. This is overcome by linking the data in an SQL with the same identifier. Ideally, the user would want to have all the data in one file as not everyone is familiar with using SQL as it is required for linking.

Unfamiliar column headings

The headings are in codes and to overcome this, there is a separate file that provides the definitions; however, it is more effective to use a database that provides all the necessary information in a common file.

3.3.1.3.3 Opportunities

Data Usage

Smaller geographical based databases

Developing smaller databases based on the different geographical areas will result in a more comprehensible database because several columns will be omitted resulting in a single file for each database. This will result in easier navigation and lead to the use of familiar languages, codes and classification used for that area, resulting in a wider usage of the database.

3.3.1.3.4 Threats

Data Usage

First time users could be deterred

The use of SQL language, heading codes, TSV file formats may be unfamiliar to first time users and the need to consult various documents for clarity may be time consuming and results in a limited number of users.

Lack of quality control

The accuracy of the database cannot be measured by first time users due to units of measurements recorded in other files.

3.3.1.4 ARC-ISCW

Table 3.5 gives the SWOT analysis of the ARC-ISCW database, while discussion of the different points will follow in the text. Based on the analysis of the ARC-ISCW database and criteria set out, the following SWOT analysis was used to propose a structure for the NWU soil database.

Table 3-5: The SWOT analysis of the ARC-ISCW database

S	**W**
• Database stored in MS Access. • Database available to MS Excel • PDF printouts for specific soil profiles available. • Availability of a User Manual. • Units of measurements included. • Availability of data validation tests. • Necessary requirements for data entry provided.	• Database not readily available. • Description documents separate from database. • No analytical methods not readily presented in the table. • Specialized skills required to extract and query data from MS access.
O	**T**
• Data availability would increase usage	• Database not used due to being unavailable • Resignation of skilled MS Access user will result in the database being unavailable.

3.3.1.4.1 Strengths

Legibility

Database stored in MS Access

The ARC-ISWC uses a database management system such as MS Access to store soil data. MS Access enables the user to perform queries to acquire specific information from the database and to analyse the relationship between the recorded soil parameters. Furthermore, the data entry program is easy to use as it provides drop down menus to enter the relevant information required for each soil profile, this information is automatically recorded in the MS Access database. The program provides different options such as viewing or printing existing data, new registration of soil profiles, lab data requests and/or updating of existing data.

Database available to MS Excel

The database can be opened and extracted into MS Excel, which is a program accepted as mastered by potential users of the database. MS Excel presents the soil data in the database in a comprehensible manner, where a soil profile is recorded with accompanying horizons in sequential rows while the specific soil properties are recorded in designated columns. Furthermore, data in Excel can be analysed and manipulated as desired without changing the structure of the original database.

Data could be presented as PDF printouts per soil profile

When data from MS Access is requested, the option to obtain soil data by printing PDF documents is available, which could be useful in certain situations, such as teaching and soil profile discussion. The PDF documents presents data in a table format, this results in easily comprehensible soil information. This soil data was clearly recorded, with each soil attribute in designated columns and clearly described as headings and the attribute features described as sub-headings in the table. Furthermore, the data was divided into columns based on the type of soil information.

Data usage

Data entries and queries in MS Access

MS Access enables the user to perform queries to acquire specific information from the database and to analyse the relationship between the recorded soil parameters. Furthermore, the data entry program is easy to use as it provides drop down menus to enter the relevant information required for each soil profile, this information is automatically recorded in the MS Access database. The program provides different options such as viewing or printing existing data, new registration of soil profiles, lab data requests and/or updating of existing data.

Availability of a User-Manual

There is a User Manual document outlining the necessary steps for processes involved in data entries, queries and data acquisition. Additionally, the User Manual provides information on the minimum data requirements for each soil profile.

Data quality

Units of measurements included

There is a unit of measurement for each soil attribute presented on the PDF printouts, resulting in easy comprehensibility and the performance of quality control to check for accuracy of the data recorded.

Availability of data validation tests

The program provides various data validation tests that can be run on the recorded soil attributes to ensure data accuracy.

Standardization of data

Necessary requirements for data entry provided

The User Manual provides information on the necessary information required before data entry. These includes: a soil classification system, horizon definition, geographic location and map sheets. This information improves the process of data standardization during data unification.

3.3.1.4.2 Weaknesses

Accessibility

Database not readily available

The database is not readily available online due to lack of funding from either the ARC or central government. The main reason for this is therefore the view of ARC that the database (and its contents/products) is a valuable source of income for the organisation and cannot therefore be made freely available at this time. Contact via email had to be made to the provider to acquire soil data and information about the original structure of the database This can sometimes be time-consuming because obtaining this information is dependent on the time it takes for the provider to reply to the request and other inquiries which may follow.

Data quality

Description documents separated from database

The description documents providing information about the soil attributes and the User Manual providing information on how the recorded data can be used are provided. However, these are saved in separate documents, and the information in these documents is not available in the database. As a result, the user may have to open multiple documents and the database at the same time, which may cause confusion and may be time consuming.

Data Usage

Specialized skills required to extract and query data from MS access.

Although MS Access is an effective program to use for creating databases, not everyone is familiar with the functions involved in data queries. The effective use of these queries requires the user to partake in courses and/tuitions in MS Access for a clear understanding of the different functions. Furthermore, the ARC-ISCW has a limited number of people with MS Access skills, which causes data availability delays when these people are not available.

Standardisation of data

No analytical methods not readily recorded in the table

Methods used to measure the acquired soil properties and metadata such as analytical methods are not readily presented in the same table with the measured soil properties. However, this information is recorded in a separate document, namely the print-out table of the analytical data per profile.

3.3.1.4.3 Opportunities

Accessibility

Data Availability would increase usage

Availability of this database would increase the opportunity for farmers, corporations, and researchers to readily have soil data that can help them with information required for productivity and sustainability, resulting in the wider usage of the database.

3.3.1.4.4 Threats

Accessibility

Database not used due to being unavailable

The ARC-ISCW is created for the provider’s use and is not readily available to the public. This database is not free, it must be bought so it can be acquired, therefore limiting database usage.

Resignation of skilled MS Access user

There is a limited number of skilled MS Access user working at the ARC-ISCW, this could result in the unavailability of the database if the user resigns from the ARC-ISCW.

3.3.2 Implications for the NWU soil database

The NWU soil database was developed for the purpose of recording, storing, managing, and sharing soil point data for carrying out collaborative research, national scale analytics and statistical analysis. The aim of this research was to analyse different soil databases with the purpose of finding a suitable structure for the NWU soil database.

3.3.2.1 Legibility

The strengths observed to ensure database legibility of the WoSIS database include the ease to import the different files making up the database into an SQL and statistical software to be recorded in a common file to allow for the handling and querying of the data. The availability of a link table allows the ease to navigate through the WoSIS database. The strengths observed to ensure database legibility for the ARC-ISCW database include the record of data in a table format and the option to open and extract the database into MS Excel resulting in easy comprehensibility of recorded soil data. The weaknesses minimizing the legibility of the WoSIS database includes recording of a large volume of data into the database resulting in a clustered presentation of data.

The strengths were incorporated in the NWU structure and weakness negated such that the NWU soil database was developed in an excel spreadsheet with a total of six table headings displaying soil information with accompanying soil attributes, units, and methods of measurements. The table format improved data presentation and comprehensibility.

3.3.2.2 Data quality

The strengths observed to ensure data quality of the WoSIS database includes the availability of articles, reports and documents proving information on the database. The strengths ensuring data quality for the ARC-ISCW database include the availability of units of measurements for the recorded soil attributes and data validation tests to ensure data accuracy. The weaknesses observed to minimise data quality of the ARC-ISCW database includes the separation of description documents proving information about the database and the recorded data from the main database.

The strengths were incorporated, and weaknesses negated in the NWU database structure such that the database is created in a workbook in MS Excel that consists of six worksheets one of which is used to describe all the attributes, units and methods of measurements, symbols, and a reference list of the source data. Furthermore, each recorded attribute is accompanied by a unit of measurement displayed in the same column.

3.3.2.3 Accessibility

The strengths observed to ensure database accessibility of the WoSIS database, include the fact that the database is readily available online. The weakness limiting the accessibility of the ARC-ISCW database is that the database is not readily available online, instead it needs to be bought from the organisation. The strengths were incorporated, and weakness negated in the NWU structure in that the database will be made freely available upon request from the NWU.

3.3.2.4 Data usage

The weaknesses limiting the data usage of the WoSIS soil database includes the use of different files making up the database saved in TSV documents which open in notepad, resulting with difficulties in reading and editing. This limits the number of users because not everyone is familiar with this format. Furthermore, the database is spread over four attribute files, although these files can be joined using the SQL software, not everyone is familiar with this software, therefore liming data usage. The database uses unfamiliar column headings and codes for the soil attributes, although there is a file proving definitions of the attributes, it is more effective to provide all the necessary information about the database in a common file. The weaknesses limiting data usage for the ARC-ISCW includes the use of MS Access to store data, there is only a limited number of people who are familiar with the functions of the software, therefore, specialized skills are required to extract and query data from MS Access. These weaknesses were negated in the NWU soil database structure such that, the database was created in MS Excel in table format making reading and editing easy without distorting the original structure of the database. All the soil profiles, units, and methods of measurements were recorded in the same columns. Furthermore, the names of the soil attributes, units and methods of measurements used in the NWU soil database are commonly used in Soil Sciences throughout South Africa.

Opportunities for increasing data usage for the WoSIS database include the development of smaller databases for different geographical areas, this will result in more comprehensible data because the data will be less clustered. This will also improve navigating through the database and use of familiar languages, codes and classifications used in specific areas. The threats observed for the WoSIS database include deterring first time users of the database due to unfamiliar document formats, SQL, and codes used for the database. This results with the need for first time users to consult various documents for clarity which can be time consuming. The threats observed for the ARC-ISCW database include limited use of the database due to its inaccessibility and a scarcity of skilled MS Access users. To incorporate the opportunities of the WoSIS database, the NWU soil database is used to record and store soil data for the geographical space of South Africa leading to the use of common languages, codes, and classification. Therefore, increasing the data usage of the NWU soil database. The threats observed in the databases were negated such that NWU soil database was made easily comprehensible and user-friendly so that anyone will be able to use. Furthermore, data recorded in the soil database can be easily imported and exported into other software using the “copy” and “paste” options.

3.3.2.5 Standardisation

The strengths observed to ensure standardisation of data in the WoSIS database include the availability of method description in the same table and column as the soil attributes. The lack of units of measurements for soil properties is one of the weaknesses limiting standardization in the WoSIS database and the lack of methods of measurement representation in the same table as recorded soil properties is one of the weaknesses limiting standardization in the ARC-ISCW database. The strengths mentioned were incorporated and weaknesses negated such that as mentioned above, the NWU soil database structure was developed in a way that the soil attribute, the units, and methods of measurements are recorded in the same column.

3.3.3 Proposed structure of the NWU soil database

Table 3.6 indicates the complete structure of the NWU database with soil data recorded in Excel workbook. All the collected data was recorded on a single worksheet and the different soil properties were further divided into separate worksheets for the users who would like to analyse a specific soil property. The first worksheet has one table with six primary table headings. Each column represents a different soil property, while the subheadings indicate the attribute features with the units and methods of measurements used to analyse the attribute. Each soil profile was recorded in a series of row, with different horizons recorded in sequential rows. The soil attribute terms used in the database are description names usually used in Soil Sciences and the description of the terms are provided in the 'Attribute Description' worksheet.

Table 3-6: Sequential presentation of the complete NWU soil database structure, with soil profile data recorded in a single worksheet.

1		PROFILE ID						LANDFORM AND TOPOGRAPHY									
2	Profile No.	Profile cd.	DD Coordinates		Name of provider	Lab name/code	Date	Slope%	Aspect	Elevation (m)	Soil Form	Soil Family	Soil Class	Master Horizon	Diagnostic Horizon	Depth (mm)	Auger Refusal
3			X	Y													
4																	
5	1	1718	27,56667	-27,0167	ARC-ISCW	-	1921/02/02	1	South	1356	Glencoe		1977 1st ed.	O		270	
6	1	1718	27,56667	-27,0167	ARC-ISCW	C4362	1921/02/02	1	South	1356	Glencoe		1977 1st ed.	A1		550	
7	1	1718	27,56667	-27,0167	ARC-ISCW	C4363	1921/02/02	1	South	1356	Glencoe		1977 1st ed.	B21	Yellow-brown apedal	1000	
8	1	1718	27,56667	-27,0167	ARC-ISCW	C4364	1921/02/02	1	South	1356	Glencoe		1977 1st ed.	B22		1100	not reached

1															
2	Soil Texture							Description							
3	Clay %	Sand %				Silt %		Colour (dry)	Colour (wet)	Colour cd. (dry)	Colour cd. (wet)	Structure	Bulk density (g/cm3)	Texture Class	Consistency
4		coSand	meSand	fiSand	vfiSand	coSilt	fSilt								
5															
6	5,70	1,40	33,30	57,60					dark yellowish brown		10YR4/4	massive		fine sand	friable
7	8,80	6,00	30,90	49,00					yellowish brown		10YR5/6	single grain		loamy medium sand	friable
8	9,00	4,40	28,80	52,90										loamy fine sand	

1			CHEMICAL PROPERTIES																							
2			NH4OAc Cations (cmol(+)/kg soil)				Mehlich 3 Cations (mg/kg)				di-ammonium EDTA Micro Nutrients (mg/kg)					DTPA Micro Nutrients (mg/kg)					Mehlich 3 Micro Nutrients (mg/kg)					
3	Transition	Other comments	Na	K	Ca	Mg	Na	K	Ca	Mg	Zn	Mn	Cu	Co	B	Zn	Mn	Cu	Co	B	Zn	Mn	Cu	Co	B
4																									
5																									
6			0,00	0,10	4,40	0,10					0,09	0,80	0,65	0,14	0,47										
7	abrupt wavy	fine <2-6mm sesquioxide concret	0,00	0,10	0,20	0,10					0,44	0,10	0,69	0,03	0,32										
8		ferricrete	0,00	0,00	1,90	0,10					0,18	0,30	0,78	0,15	0,32										

1																	
2	Phosphorus Status					KH_2PO_4 P-Soption (%)	pH		CBD			CEC (cmol/kg)	EC Saturated Paste (mS/m)	EC (mS/m)	WBM-Black OC %	Dry combustion OC%	OC %
3	ISFEI P(mg/kg)	Olsen P(mg/kg)	Mehlich 3 P(mg/kg)	Bray 1 P (mg/kg)	Bray 2 P (mg/kg)		H2O	KCl	Fe%	Al%	Mn%						
4																	
5																	
6	0,90					27,04	5,10		0,37	0,04	0,00	1,70			0,30		
7	0,40					63,04	4,80		0,54	0,07	0,00	2,30			0,20		
8	0,70					87,49	5,80		2,11	0,22	0,00	3,00			0,20		

1	HYDROLOGICAL PROPERTIES																						GEOLOGICAL PROPERTIES	
2	Water retention %				Water retention (mm)												AWR %	Hydraulic conductivity (mm/ hr^{-1})					Saturated Hydraulic Conductivity (Ks) (	XRD Mineralogy	Parent material and/or Supergroup
3	-33kPa	-80kPa	-500kPa	-1500kPa	0	38	50	100	200	400	600	800	1000	8000	10000	15000		0mm	30mm	60mm	80mm	150mm		Mineral symbols (μm)	
4																									
5																									
6	3,70	3,20	2,60	2,40													0,20							Qz, Fsl, Tcl, Kt,Tc, Mi, Ch, Vm	
7	5,50	4,70	3,80	3,60													0,10							Qz, Fsl, Tcl, Kt,Tc, Mi, Ch	
8	8,60	6,80	5,20	4,80													0,10							Qz, Tc, Fsl, Kt, Mi, Ch	

The soil attributes used to identify the soil profile (profile id.) are shown under the first table heading: profile no. as defined by the database, and profile cd. (code) as defined by the source. Furthermore, the origin of each soil is characterised by the co-ordinates-longitude (Y) and latitude (X) in the format of Decimal Degrees, the source of the data and lab name and/or code where the soil analysis was conducted and the date for soil collection and /or analysis (Table 3.7).

Table 3-7: Excel worksheet showing the first table heading and profile identification attributes as recorded in the structure of the NWU soil database.

1	PROFILE ID						
2	Profile No.	Profile cd.	DD Coordinates		Name of provider	Lab name/code	Date
3			X	Y			
4							
5	1	1718	27,566666	-27,01667	ARC-ISCW	-	1921/02/02
6	1	1718	27,566666	-27,01667	ARC-ISCW	C4362	1921/02/02
7	1	1718	27,566666	-27,01667	ARC-ISCW	C4363	1921/02/02
8	1	1718	27,566666	-27,01667	ARC-ISCW	C4364	1921/02/02

The second table heading represents the landform and topography information of the soil profile, which serves to give an idea of the surface and arrangement of the landscape the soil profile originated from. This information is represented by the slope, aspect, and elevation of the surface (Table 3.8). The profile number and code appear with all the table headings as a ‘freeze’ pane was used to lock the first two columns to make data entry easier when the volume of data increases. This was also done for the first two rows of the database to lock the table headings, profile identification attributes and soil attributes.

Table 3-8: Excel worksheet showing the second table heading and landform and topography soil attributes as recorded in the structure if the NWU soil database.

1			LANDFORM AND TOPOGRAPHY		
2	Profile No.	Profile cd.	Slope%	Aspect	Elevation (m)
3					
4	1	1718	1	South	1356
5	1	1718	1	South	1356
6	1	1718	1	South	1356
7	1	1718	1	South	1356

Morphological and physical properties of the soil profile are shown under the third table. The soil classification aspects include soil form, soil family, soil classification system, master horizons, and diagnostic horizons. The technical terms used to describe these features are explained in the book “Soil classification: A Binomial System for South Africa” (MacVicar *et al.*, 1977). This column also contains information on the depth of each soil horizon tested and the auger refusal status. In addition, the soil texture, colour (wet and dry), colour code (wet/dry)-as determined by the Munsell system, structure, bulk density (g/cm^3) texture class, consistency, transition, and other comments on the physical characteristics of the soil profile are recorded in this column (Table 3.9).

Table 3-9: Excel worksheet showing the third table heading and morphological and physical descriptive soil attributes as recorded in the structure if the NWU soil database.

1	SOIL MORPHOLOGICAL AND PHYSICAL PROPERTIES								
2	Profile No.	Profile cd.	Soil Form	Soil Family	Soil Class	Master Horizon	Diagnostic Horizon	Depth (mm)	Auger Refusal
3									
4									
5	1	1718	Glencoe		1977 1st ed.	O		270	
6	1	1718	Glencoe		1977 1st ed.	A1		550	
7	1	1718	Glencoe		1977 1st ed.	B21	Yellow-brown apedal	1000	
8	1	1718	Glencoe		1977 1st ed.	B22		1100	not reached

1													
2	Profile No.	Profile cd.	Soil Texture										
3			Clay %	Sand %				Silt %		Colour (dry)	Colour (wet)	Colour cd. (dry)	Colour cd. (wet)
4				coSand	meSand	fiSand	vfiSand	coSilt	fSilt				
5	1	1718											
6	1	1718	5,70	1,40	33,30	57,60					dark yellowish brown		10YR4/4
7	1	1718	8,80	6,00	30,90	49,00					yellowish brown		10YR5/6
8	1	1718	9,00	4,40	28,80	52,90							

1								
2	Profile No.	Profile cd.	Description					
3			Structure	Bulk density (g/cm3)	Texture Class	Consistency	Transition	Other comments
4								
5	1	1718						
6	1	1718	massive		fine sand	friable		
7	1	1718	single grain		loamy medium sand	friable	abrupt wavy	fine <2-6mm sesquioxide concret
8	1	1718			loamy fine sand			ferricrete

The fourth table was used to record the chemical properties of the soil profiles including cations (Na, K, Ca, Mg) which can be recorded in one of the two columns for the different methods including Ammonium Acetate (cmol (+)/ kg) and Mehlich 3 (mg/kg). Micronutrients (Zn, Mn, Cu, Co, B) could be recorded in one of three columns of different methods including Di-ammonium EDTA and Diethylenetriamine pentaacetate and Mehlich 3 measured in mg/kg. Phosphorus status also displayed by five methods of measurements, ISFEI, Olsen, Mehlich 3 methods and the Bray Test 1 and 2, measured in mg/kg. P-sorption measured with KH_2PO_4 in percentage (%). The pH was displayed by two methods of measurements including H_2O and KCL. Extractable irons (Fe, Al, Mn) measured with Citrate-bicarbonate-dithionite (CBD), Cation Exchange Capacity (CEC) measured with Ammonium Acetate in cmol (+)/kg soil a, EC measured with saturated paste and EC with no specified method of measurement-both measured in mS/m, Organic C displayed by two methods, Walkley-Black and Dry Combustion method and an additional column representing Organic Carbon values with no specified method of measurements. Organic carbon is measured in percentage values in all the methods. The above-mentioned soil attributes are utilized to provide information on the chemical status of the soil profiles (Table 3.10).

Table 3-10: Excel worksheet showing the fourth table heading and chemical soil attributes as recorded in the structure if the NWU soil database.

1			CHEMICAL PROPERTIES												
2	Profile No.	Profile cd.	NH4OAc Cations (cmol(+)/kg soil)				Mehlich 3 Cations (mg/kg)				di-ammonium EDTA Micro Nutrients (mg/kg)				
3			Na	K	Ca	Mg	Na	K	Ca	Mg	Zn	Mn	Cu	Co	B
4															
5	1	1718													
6	1	1718	0,00	0,10	4,40	0,10					0,09	0,80	0,65	0,14	0,47
7	1	1718	0,00	0,10	0,20	0,10					0,44	0,10	0,69	0,03	0,32
8	1	1718	0,00	0,00	1,90	0,10					0,18	0,30	0,78	0,15	0,32

1																		
2	Profile No.	Profile cd.	DTPA Micro Nutrients (mg/kg)					Mehlich 3 Micro Nutrients (mg/kg)					Phosphorus Status				
3			Zn	Mn	Cu	Co	B	Zn	Mn	Cu	Co	B	ISFEI P(mg/kg)	Olsen P(mg/kg)	Mehlich3 P(mg/kg)	Bray1 P (mg/kg)	Bray2 P (mg/kg)
4																	
5	1	1718															
6	1	1718											0,90				
7	1	1718											0,40				
8	1	1718											0,70				

1														
2	Profile No.	Profile cd.	KH_2PO_4 P-Soption (%)	pH		CBD			CEC (cmol(+)/kg)	EC (ms/m) Saturated Paste	EC (mS/m)	WBM-Black OC %	Dry combustion OC%	OC %
3				H2O	KCl	Fe%	Al%	Mn%						
4														
5	1	1718												
6	1	1718	27,04	5,10		0,37	0,04	0,00	1,70			0,30		
7	1	1718	63,04	4,80		0,54	0,07	0,00	2,30			0,20		
8	1	1718	87,49	5,80		2,11	0,22	0,00	3,00			0,20		

The fifth column represents information on the hydrological properties and this attribute is defined by features such as water retention measured in percentage at different atmospheric pressures (kPa) and measured in mm at different depths. Relative soil saturation (AWR) measured in percentage, hydraulic conductivity measured at different depths with the unit measurement mm/hr-[1] and IN-SITU Saturated Hydraulic Conductivity measured in mm/ hr-[1] are also displayed in this column (Table 3.11).

Table 3-11: Excel worksheet showing the sixth table heading and hydrological soil attributes as recorded in the structure if the NWU soil database.

1			HYDROLOGICAL PROPERTIES																							
2	Profile No.	Profile cd.	Water retention (%)				Water retention (mm)												AWR %	Hydraulic conductivity (mm/ hr^{-1}					IN-SITU Saturated Hydraulic Conductivity (Ks) (mm/hr1)
3			-33kPa	-80kPa	-500kPa	-1500kPa	0	38	50	100	200	400	600	800	1000	8000	10000	15000		0mm	30mm	60mm	80mm	150mm	
4																									
5	1	1718																							
6	1	1718	3,70	3,20	2,60	2,40													0,20						
7	1	1718	5,50	4,70	3,80	3,60													0,10						
8	1	1718	8,60	6,80	5,20	4,80													0,10						

The final column represents information on the geological properties of each soil profile, identifying the minerals encountered in the soil which were determined using the XRD method and a column displaying the parent material from which the soil was developed and/or Supergroup underlying the analysed soil profile (Table 3.12).

Table 3-12: Excel worksheet showing the sixth table heading and geological soil attributes as recorded in the structure if the NWU soil database.

1			GEOLOGICAL PROPERTIES	
2	Profile No.	Profile cd.	XRD Mineralogy	Parent material and/or Supergroup
3			Mineral symbols (µm)	
4				
5	1	1718		
6	1	1718	Qz, Fsl, Tcl, Kt,Tc, Mi, Ch, Vm	
7	1	1718	Qz, Fsl, Tcl, Kt,Tc, Mi, Ch	
8	1	1718	Qz, Tc, Fsl, Kt, Mi, Ch	

The Attribute Column worksheet consists of two tables, the first table describes the soil attributes, the methods of measurements and the units of measurements. The second table describes the mineral symbols and the full name of the minerals. The database structure was created in such a way that the tables from the NWU soil database sheet and the Attribute Description sheet are interlinked. This was accomplished by constructing hyperlinks for all the attribute features, so that when a feature is selected from the NWU soil database sheet, the description table in the Attribute Description sheet can be readily accessed. Thus, if the user does not understand the meaning of a soil attribute and needs additional information about that specific attribute, they may quickly access description sheet by selecting the attribute feature from the table. The hyperlink from the Attribute Description sheet goes back to the attribute features in the main table.

The Morphological and Physical Properties, Chemical Properties, Hydrological Properties, and Geological Properties worksheets were established to provide soil information on a single attribute so that if the user wishes to utilize a specific soil attribute data (e.g., chemical soil analysis), they may do so without being distracted by the other recorded data.

3.4 Conclusion

The structure of the NWU soil database was established through the SWOT analysis of the WoSIS and ARC-ISCW databases, where the strengths, weaknesses, opportunities, and threats of each were evaluated.

The SWOT analysis results indicated that the strengths that promoted the legibility of the WoSIS and ARC-ISCW database include the ease to import data into a common database and other software and data presentation using table formats to increase data comprehensibility. The weaknesses that limited data legibility include the impact of recording large volumes of data into one database as the data appears cluttered, limiting comprehensibility.

The strengths that promoted data quality include the availability of documents providing important information about the database. Furthermore, the data quality of a database is dependent on the presence of information including units and methods of measurements used for each soil attribute. The weaknesses that limited data quality of the database includes lack of description documents.

The strengths that promoted the data usage include the ease to navigate, read, edit, analyse, import, query, perform quality control and compare the data in the database. Furthermore, the weaknesses that limit data usage includes the ease to understand the software used to store and record data, the language, codes, and classifications used in the database, plays a vital role on how useful the database will be to people and who is able to use the database.

The strengths that promoted accessibility of the database include the ease to obtain the database. The strengths that promoted the standardisation of the databases include the availability of units and methods of measurements used for each soil attribute to promote quality control and the weaknesses that limited standardisation is the lack thereof.

The strengths from the SWOT analysis results of the WoSIS and ARC-ISCW were incorporated for the NWU soil database structure development and weaknesses negated. Therefore, the NWU soil database structure was developed in Excel software with all the soil attributes, units and methods of measurements recorded in a common worksheet. The database consists of six different worksheets one of which consists of all the description information for the database and its contents. The Excel spreadsheets promotes ease to navigate, read, edit, analyse, import, query, perform quality control and compare. Furthermore, the database is used to record soil data in South Africa using language, terminology, codes, and classification as used by South African Soil Scientists and researchers. The NWU soil database will be made available upon request from the NWU.

3.5 References

Fern‘ndez, R.N. & Rusinkiewicz, M. 1993. A conceptual design of a soil database for a geographical information system. I*nternational Journal of Geographic Information Systems*, 7(6): 525-539.

Grealish, G., King, P., Omar, S. & Roy, W. 2004. Geographic information system and database for the soil survey for the State of Kuwait-design and outputs. *Kuwait Journal of Science and Engineering*, 31 (1);135-148.

Macvicar, C.N., De Villiers, J.M., Loxton, R.F., Verster, E., Lambrechts, J.J.N., Merryweather, F.R., Le Roux, J., Van Rooyen, T.H. & Harmse, V.M. 1977. Soil Classification: A Binomial System for South Africa. 1st edition, Pretoria: The Department of Agricultural technical services republic of South Africa.

Melikhova, E.V., Rogachev, A.F. & Skiter, N.N. 2019. Information System and Database for Simulation Irrigated Crop Growing. Ubiquitous Computing and the Internet of Things: Prerequisite for the Development of ICT, p1185-1191.

Ribeiro, E., Batjes, N.H. & Oostrum, A.J.M. 2020. World Soil Information Service (WoSIS)- Towards the standardization and harmonisation of world soil profile data. Procedure manual 2020, Report 2020/01, ISRIC-World Soil Information. http://dx.doi.org/10.17027/isric-wdc-2020-01

Wösten, J.H.M., Lilly, A., Nemes, A. & Le Bas, C. 1999 Development and use of a database of hydraulic properties of European soils. *Geoderma*, 90: 169-185. https://doi.org/10.1016/S0016-7061(98)00132-3

CHAPTER 4 DATA POPULATION IN THE NWU SOIL DATABASE

4.1 Introduction

After determining the structure of the NWU soil database, it needed to be populated. As alluded to in the literature review there are various sources of soil data available, however, this data is not available for public use (Romero *et al*., 2012:163). Furthermore, data was collected for different purposes, thus each dataset was recorded in different formats, measured with different methods, and expressed using a variation of units of measurements (Dobos *et al*., 2010:13). Therefore, an initiative was taken to collect soil data from different sources and populate it into the NWU soil database. This process requires harmonization and standardization of data to incorporate all the soil data with different methods and units of measurements into the NWU soil database.

Data harmonisation allows spatial and attribute data to be integrated into a single system to allow a comparable view of data from different sources (Sulaeman *et al*., 2013:78). Data standardization involves bringing data into a common format to eliminate variation in classification, terminology, and measurements, for better data quality and interpretation (Hortensius & Nortcliff, 1991:165). In this chapter the procedures followed for harmonization and standardization of soil data from various sources to be incorporated into the NWU soil data was discussed. This was done to provide good quality soil data representing the geographical extent space of South Africa.

4.2 Materials and Methods

4.2.1 Sources

Soil data in the NWU soil database comes from a variety of sources and is recorded in a variety of formats depending on the aim of data collection. Some of the data was transcribed into reports in text and tabular format, which were saved as Word or PDF documents. Other information was saved as tabular datasets in excel or csv documents. The data sources that were included in the NWU soil database include:

4.2.1.1 Academic transcripts:

- An integrated approach for the delineation of arable land and its cropping suitability under variable soil and climatic conditions in the Nkonkobe municipality, Eastern Cape, South Africa (Manyevere, 2014).
- Impacts of soil information on process based hydrological modelling in the upper Goukou catchment, South Africa (Smit, 2021)
- Investigating the suitability of Land Types for hydrological modelling in the mountain region of Hessequa South Africa (Malan, 2016).

4.2.1.2 Reports

- Irrigation Suitability Report for the farm Backhouse, Douglas, Northern Cape Province for Wim van Bergen (Van Zijl & Du Plessis, 2016).
- Irrigation Suitability Report for a portion of land identified by Mahoebe Eiendomme, Prieska, Northern Cape Province for Henry Coetzee (Le Roux *et al*., 2016).
- Irrigation Suitability Report for the Botha Farm, Prieska, Northern Cape Province for Jan-Philip Botha (Van Zijl & Du Plessis, 2017).
- Irrigation Suitability Report on Kheis 291 Groblershoop Northern Cape Province for Wikus Snyman (Van Zijl, 2015).
- Irrigation Suitability Report to cultivate Pecan nuts on a portion of Spitskop farm, Prieska, Northern Cape Province for Johan Gous (Van Zijl & Du Plessis, 2017).
- Irrigation Suitability Report on Fam 484 near Kolomela Mine, Postmasburg Northern Cape Province for Umhlaba Consulting (Van Zijl & Du Plessis, 2015).
- Irrigation Suitability Report for the Fieldsview Irrigation Project, Riverton, Northern Cape Province for Milnex (Van Zijl *et al*, 2017)
- Irrigation Suitability Report to cultivate Pecan nuts on a portion of Uitdraai farm, Prieska, Northern Cape Province for Orffer Muller (Van Zijl & Du Plessis, 2017).
- Irrigation suitability Report for the farm Witdam, near Kimberley, North Cape Province for van Deventer Familie Trust (Van Zijl & Du Plessis, 2017).
- Irrigation suitability of Bucklands portion 15 & 16, Douglas, Northern Cape (Van Tol & Marx, 2020)
- Irrigation suitability of Selene, Douglas, Northern Cape (Van Tol & Marx, 2020).
- Irrigation suitability of Willowdene, Barkely-West (Van Tol & Marx, 2021)
- Irrigation suitability report for the farm Harrisdale portion 223-Kilmorey, near Kimberley, Northern Cape Province for Turn180 (Van Tol & Le Roux, 2019).
- Hydropedological Survey Report-ABB, NIGEL (Van Tol & Le Roux, 2015).
- Hydropedological survey of the Vopak site, Port of Richards Bay, KwaZulu-Natal (Van Tol & Bouwer, 2019).
- Hydropedological survey of the Der Brochen Mine Exploration project, Limpopo Province (Van Tol & Bouwer, 2018).
- Report on the Irrigation suitability of Kameelhoek, Postmasburg, Northern Cape (Van Tol & Marx, 2021)
- WAD 510 (Klooffontein) Irrigation suitability soil survey (Van Tol & Marx, 2020).

4.2.1.3 Research Reports

- Characterisation and mapping of the erodibility of soils in selected land types in support of the Ntabalanga and Laleni ecological infrastructure project (NLEIP) (Van Tot et al., 2018).
- Hydrology of South African Soils and Hillslopes (HOSASH) (Le Roux *et al*., 2015).
- Hydrological Processes Research: Experiments and measurements of soil hydraulic characteristics (Lorentz *et al*., 2001).

4.2.1.4 Datasets

- ARC-ISCW (ARC-ISCW, 2021).
- Database of hydraulic properties (Van Tol, 2021).

4.2.2 Data Population

Excel tables were used for data entry and collation, ensuring pragmatism and speed. Unique soil profile IDs were used when recording soil data to avoid duplication, these IDs are used to reference the original profile IDs used in the different data sources. Prior to the population process, each dataset was analysed as to observe the units and methods of measurements used to record the data in the designated columns. If the measuring units were different from the standard measurement units used in the NWU soil database, a conversion process was carried out to ensure standardisation. Datasets that were recorded in PDF documents were copied manually into the database, while datasets which were saved in excel documents were copied using the "copy" and "paste" options. The former was time consuming and prone to human error while the latter improved time effectiveness and accuracy.

4.2.3 Data legibility

The criteria used for the inclusion of soil data into the NWU soil database involves the presentation of the name of the provider and at least one soil attribute. Co-ordinates were not included as a prerequisite because data collected from various sources did not all contain co-ordinates, however the remaining data contained important information that can be useful for research such as the creation of pedotransfer functions. Furthermore, minimum data requirements for soil profiles include information that aids to soil classification, such as soil colour, texture, and/or structure. The soil attributes recorded in the NWU soil database include the profile code, geographic co-ordinates, name of provider, lab name/code, slope %, aspect, elevation, soil form, soil family, soil class, master horizon, diagnostic horizon, depth, soil texture, colour (dry and wet), colour code (dry and wet), structure, bulk density, texture class, consistency, transition, other comments, Ammonium acetate-cations, Mehlich 3-cations, di-ammonium EDTA M-micro nutrients, DTPA-micro nutrients, Mehlich 3- micro nutrients, Phosphorus status,, P-sorption, pH (H_2O , KCl), CBD (Fe, Al, Mn), CEC, EC, Organic Carbon, water retention, Relative Saturation (AWR), hydraulic conductivity, saturated hydraulic conductivity, XRD-mineralogy and parent material and/ or

Supergroup. Furthermore, the units and methods of measurement are also essential attributes that affect the quality of the data. These attributes provide measures that enable the database user to identify variables that may comprise the quality of specific data, hence determining its eligibility for use. The use of these soil attributes as quality indicators guarantees that objective approaches are used to evaluate the data within the database.

4.2.4 Standardization

4.2.4.1 Methods of standardization carried out

4.2.4.1.1 Conversion of DSM to DD Calculation:

Data in geographical degrees, minutes, seconds (DMS) had to be converted to decimal degrees (DD). Equation 1 was used for this:

$$DD = D + (M/60) + (S/3600). \quad (1)$$

4.2.4.1.2 Conversion of degrees, decimal minutes to DD Calculation (Equation 2):

$$DD= D+M/60. \quad (2)$$

4.2.4.1.3 Calculation of geographical co-ordinates:

Geographical co-ordinates were not recorded in some of the DSA reports; however, these reports contained shapefiles, and latitude and longitude values had to be calculated using the ARC-GIS tools recorded in the shapefiles.

4.2.4.1.4 Conversion of cation values from mg/kg to cmolc/kg (Equation 3-6):

$$Ca_{cmolc/kg} = Ca/_{mg/kg} \text{ x } 200 \quad (3)$$

$$Mg_{cmolc/kg} = Mg/_{mg/kg} \text{ x } 120 \quad (4)$$

$$K_{cmolc/kg} = K/_{mg/kg} \text{ x } 390 \quad (5)$$

$$Na_{cmolc/kg} = Na/_{mg/kg} \text{ x } 230 \quad (6)$$

4.2.4.1.5 Basic conversion of units of measurements:

Profiles and horizon depths from cm to mm (Equation 7)

$$Depth_{mm} = Depth_{cm} \text{ X } 10 \quad (7)$$

Bulk Density (BD) from various units to g/cm^{-3} (Equations 8 and 9)

$BD_{g/cm^3} = BD_{kg/m^3} / 1000$ (8) (8)

$BD_{g/cm^3} = BD_{mg/m^3} / 0.000000000$ (9) (9)

4.2.4.2 Standardization of data sources

4.2.4.2.1 PhD Thesis (Manyevere, 2014)

The depth of the soil horizons in the thesis was recorded in cm units; these values were converted to mm as the standard unit of measurements for depth in the NWU soil database. Some of the units of measurements were not provided in the table describing chemical properties, however, the methods of measurements were provided in the content of the thesis, but not incorporated in the table. Therefore, the user had to read through the thesis to gather this information.

Data presentation was done via table format making the data easily comprehensible, however, the PDF document resulted in manual data entry. On the descriptive soil data, the different layers were designated by the different depths instead of soil horizons, thus the soil attributes were recorded per soil form instead of horizon. During data population, it was encountered that there were two families within the Hutton soil form, Hutton (dystrophic) and Hutton (eutrophic). This data was recorded without specifications of the type of Hutton soil form in the soil description. This resulted in difficulties in determining the soil family, thus excluded from population.

4.2.4.2.2 Goukou Data (Smit, 2021)

The co-ordinates from the database were converted using excel functions from DMS to DD, steps are displayed in Annexure B. Soil depth was also recorded in cm, these units were converted to mm. Due to the data being saved in different workbooks, it was time consuming to go through different workbooks to gather information for a single soil profile.

4.2.4.2.3 Hydrological Modelling Data (Malan, 2016)

Due to a lack of chemical and hydrological properties, there was no standardisation necessary. The depth was recorded in the same unit of measurement as used in the NWU soil database (mm). Furthermore, there were no challenges encountered during data population, as the database was saved in Excel and the information was copied and pasted into the NWU soil database.

4.2.4.2.4 ARC-ISCW (ARC-ISCW, 2021)

During data population of the ARC-ISCW soil database, latitude and longitude values were converted from Degrees Minute Seconds (DMS) to Decimal Degrees (DD) using the method as described in the standardisation section. This calculation method was used because the data received was saved in PDF method, data could not be copied or added-therefore the data was added manually. All the data points provided by the ARC-ISCW met the requirements for inclusion in the NWU database.

4.2.4.2.5 Digital Soil Africa (DSA)

Due to a variety of DSA reports from different soil observers, data is recorded using different formats and measurements. Some reports recorded cations units of measurements as cmol (+)/kg while others recorded the values in mg/kg. Some of the reports did not have information on the co-ordinates, however, there were shapefiles in the datasets. Some of these shapefiles displayed the latitude and longitude values when recorded in ARC-GIS, and for other reports the latitude and longitude had to be calculated using ARC-GIS tools. The methods of measurement were not specified in other reports, this data was recorded in designated columns created in NWU soil database to record properties with no method specification. Although there were some reports in which the soil colour was recorded, however, the state (wet/ dry) of the soil colour was not mentioned. This information was added to the 'other comments' column in the database.

4.2.4.2.6 NLEIP Report (Van Tol *et al*., 2018)

Not all the soil attributes are accompanied by a unit of measurement, this information is also not available in the report. However, the report provides the methods of measurements for the chemical properties. Missing information about the units and methods of measurements slowed down the process of data entry as this information was searched from the report. Furthermore, the presentation of data in a PDF document halts the process of copying and pasting data from one dataset to the next. Manual data entry is time consuming and prone to human error.

4.2.4.2.7 HOSASH (Le Roux *et al*., 2015)

During data population of the soil data from the HOSASH database, it was discovered that the latitude and longitude values of the data was not uniform, some of the geographic co-ordinates were measured in DD, DMS and degrees decimal minutes, which led to conversion of all the coordinates to DD to maintain a uniform measurement and meet the measurements standard set for NWU soil database. The soil morphological and analytical data was recorded in separate worksheets resulting in difficulty of finding information for a single soil profile. Additionally, soil morphological data table was a copied image/photo which did not allow copying and pasting, the information had to be recorded manually. In some cases, data from the two worksheets did not match. Diagnostic horizon, soil form and depth from the 'soil properties and hydrology' worksheet did not match with the information recorded in the descriptive tables from the other worksheets. This resulted in the omission of such data from the NWU soil database. Although some of the described soil profiles did not have any profile no. nor ID, this data was recorded in the NWU soil database as it provided other important soil attributes.

4.2.4.2.8 Hydrological Processes report (Lorentz *et al*, 2001)

Bulk density values had to be converted from kg/m^3 to g/cm^3 as used for the standard unit of measurements for the NWU soil database. Data had to be copied manually because the report was saved in PDF format, therefore this process was time consuming and prone to human error. Furthermore, table headings were not included on each continuing page, it was time consuming going back and forth between the pages to match every column with a heading.

4.2.4.2.9 Data of hydraulic properties (Van Tol, 2021)

There was no consistency with the data in the database. The soil horizons for a specific soil profile were not recorded sequentially, corresponding soil horizons had to be searched through the database. Furthermore, for certain soil profiles a single horizon was recorded more than once because it was divided into different depths within a soil horizon.

4.3 Results and Discussion

4.3.1 PhD Thesis (Manyevere, 2014)

Soil data obtained from PhD thesis (Manyevere, 2014) was received from the University of Fort Hare, saved as a PDF document. The thesis contained a total of 25 soil profiles and 76 distinctive soil horizons. The descriptive soil data was recorded on separate sheets for each soil profile and attribute soil data was recorded in different tables, the two sets of soil data was recorded on different pages of the thesis.

The descriptive soil data sheet contained soil profile attributes including profile no., location, elevation, co-ordinates, landscape form and shape, slope position, percentage and aspect, erosion severity and type, capping, land use, effective depth, parent material, soil classification and description. Table 4.1 displays all the chemical and physical soil properties recorded in the database and the units of measurement. No hydrological properties were recorded in this dataset but there was information on the geological properties.

4.3.2 Goukou Data (Smit, 2021)

Dataset collected from the Goukou Data resulted in a total of 24 soil profiles and 57 distinctive soil horizons. The database was recorded in three different Excel workbooks. The first workbook displayed information on the co-ordinates of the soil profiles and the second workbook displayed hydrological soil properties including bulk density and saturation. The third workbook displayed morphological and physical soil properties including soil horizon, soil form, depth, texture, structure, colour code, estimated clay %, mottles, terrain morphological unit (TMU) and elevation.

4.3.3 Hydrological Modelling Data (Malan, 2016)

A total of 38 soil profiles and 109 distinctive soil horizons were recorded from the Hydrological Modelling Data dataset. All the soil attributes were recorded in a single Excel table and soil attributes for each soil profile was displayed in a single row. The database is made up of a total of four worksheets, the first worksheet was used to record all the soil profiles and soil attributes. The second and third worksheets were used to record separate information or comments and the last worksheet was used to record the soil form abbreviations and descriptions used in the database.

This database only recorded soil descriptive properties including latitude and longitude, profile no., land type, elevation, soil form and abbreviations, soil family, soil horizons, depth, colour, structure, coarse fragments, other characteristics, terrain, texture, land type soil match, and land type segmentation. No chemical, hydrological and geological soil data was recorded in the database. Therefore, the units and methods of measurements were not included.

4.3.4 ARC-ISCW (ARC-ISCW, 2021)

A sample subset of data received from ARC-ISCW, provided a total of 19 soil profiles with 60 distinctive soil horizons. The data was recorded in PDF software and the soil descriptive and attribute data was recorded in table format (figure 3.5 and table 3.3). The following soil attributes were recorded in the table: (i) soil profile description that contained information on soil form and series, soil horizons, depth, diagnostic horizon and description data; (ii) environmental attributes which incorporated surface rockiness and stoniness, wind and water erosion, vegetation, land use, water table, parent material, underlying material and weathering of underlying material and (iii) chemical, physical, hydrological and mineralogical soil properties (Table 4.2). Additionally, there is lab code for each analysis that was carried out.

4.3.5 Digital Soil Africa (DSA)

Datasets received from the DSA reports resulted with a total of 153 soil profiles and 421 distinctive soil horizons. This dataset was recorded in report format, and it is made up of a total of 18 DSA reports saved as pdf and MS Word documents. The descriptive and attribute data is recorded in tables and saved in the annexures of the report. Table 4.3 represents the chemical, physical, and hydrological properties.

4.3.6 NLEIP Report (Van Tol *et al*., 2018)

A report from the University of Fort Hare was received, which was saved as a PDF document and a total of 49 soil profiles and 109 distinctive horizons were captured from this report. The chemical, physical, hydrological, and descriptive properties were recorded in different tables and annexures.

The soil descriptive data was stored in the report includes site, GPS position, surface stones, altitude, terrain unit, slope, slope shape, aspect, micro-relief, parent material solum, geological group, soil form, soil family, diagnostic and master horizons colour, occurrence of flooding, wind and water erosion potential, vegetation/land use, water table, name of describer, date of description and weathering of underlying material. Table 4.4 is a representation of all the chemical, physical and hydrological properties. The parent material solum, geological group and mineralogy demonstrate the geological properties.

4.3.7 HOSASH (Le Roux *et al*., 2015)

A total of 51 soil profiles and 174 distinctive soil horizons were obtained from the HOSASH database. The data was recorded in Excel table, the soil descriptive and attribute data was recorded in two different worksheets. The database is made up of a total of 11 worksheets within one workbook, 10 of these contained soil descriptive data of the 10 different sampled sites and the one worksheet contained all the soil chemical and hydrological data.

The soil descriptive data that was incorporated in the database includes the profile no, map/photo no., profile geographic co-ordinates, altitude, terrain unit, slope, slope shape, aspect, microrelief parent material solum, underlying material, weather conditions, soil temperature regime, moisture regime, soil form, soil family, horizon, depth(mm), surface rockiness, occurrence of flooding, wind and water erosion, vegetation/land use, water table, name of describer, data description, weathering and alterations of underlying material and former weather conditions. Table 4.5 displays all the chemical, physical, hydrological, and mineralogical soil properties recorded in the database and the units of measurements thereof. The mineralogical properties in this case refer to the weathered and altered underlying materials.

4.3.8 Hydrological Processes report (Lorentz et al, 2001)

Soil data collected from the report resulted in a total of 93 soil profiles and 321 distinctive soil horizons. The report was received as a PDF document and soil data was recorded in a table format. The displayed soil descriptive properties included the description name and of the soil profile and the depth of each horizon. Furthermore, the table was divided into laboratory measured properties and IN-SITU measured proprieties. Table 4.6 shows the soil properties as recorded in the table.

4.3.9 Data of hydraulics properties (Van Tol, 2021)

A total of 87 soil profiles and 191 distinctive soil properties were collected from the Data of hydraulic properties database. The database consisted of a single Excel workbook with all the soil data recorded in a single table. The descriptive soil properties that were recorded include the profile no., co-ordinates, soil form, soil horizon, texture, structure, amount, and size of roots. Table 4.7 shows the chemical and hydrological data that was recorded. Methods of measurements were not specified in the database.

Table 4-1: Chemical, and physical properties and the accompanying units and methods of measurements as recorded in the PhD Thesis (Manyevere, 2014).

Soil Attribute	Units of measurement	Methods of measurement
Chemical Properties		
Acid saturation		
pH	H_2O	1:2.5 water suspension
Organic Carbon	%	Dry combustion
Phosphorus	%	Bray 1
Cations (Mg, Ca, Na, K)	cmol (+) /kg soil	1.0 M Ammonium acetate
Free (Fe)	%	NS
Exchangeable Sodium Percentage (ESP)	%	NS
Leaching status	%	NS
Physical Properties		
Particle size distribution (Cl, coSi, fiSi, coSa, meSa, fiSa, vfSa)	%	Pipette method

NS- Not Specified
Cl- Clay
coSi- Coarse Silt
fiSi- Fine Silt
coSa- Coarse Sand
meSa- Medium Sand
fiSa- Fine Sand
vfSa- Very fine Sand

Table 4-2: Chemical, physical, hydrological, and mineralogical properties and the accompanying units and methods of measurements as recorded in the ARC-ISCW database (ARC-ISCW, 2021).

Soil Attribute	Units of measurement	Methods of measurement
Chemical Properties		
Extractable Irons (Fe, Al, Mn)	%	NS
Organic Carbon (OC)	%	NS
Cation Exchange Capacity (CEC)	cmol (+)/ kg soil	NS
Extractable and Soluble Cations	cmol (+) /kg soil	1.0 Ammonium acetate
Extractable acidity	cmol (+) /kg soil	NS
pH	H_2O/KCl	1:2.5 water suspension
Micronutrients (Zn, Mn, Cu, Co, B)	mg/kg	Di-ammonium EDTA
Phosphorus Status (P)	mg/kg	ISFEI
Phosphorus Sorption	%	KH_2PO_4
Physical Properties		
Particle size distribution (coSand, meSand, fiSand, Silt, Clay)	%	NS
Hydrological Properties		
Resistivity	Ohms	NS
Electrical Conductivity (EC)	mS/m	Saturated paste
Water Retentivity (-33kPa, -80kPa, -500kPa, -1500kPa)	%	NS
Relative soil saturation (AWR)	NS	NS
MR	KPa	
Geological Properties		
Mineralogy	ɥm	XRD

NS- Not Specified
XRD- X-Ray Diffraction

Table 4-3: Chemical, and physical properties, hydrological and the accompanying units and methods of measurements as recorded in the DSA Reports.

Soil Attribute	Units of measurement	Methods of measurement
Chemical properties		
Extractable Irons (Fe, Al, Mn)	%	NS
Cation Exchange Capacity (CEC)	cmol (+)/ kg soil	NS
Extractable and Soluble Cations	cmol (+) /kg soil	1:10 Ammonium acetate/Mehlich 3
Extractable acidity	cmol (+) /kg soil	NS
pH	H_2O/KCL	1:2.5water/KCl suspension
Micronutrients (Zn, Mn, Cu, Co, B)	mg/kg	DTPA
Phosphorus Status (P)	mg/kg	Bray 1/ Bray 2
Phosphorus Sorption	%	NS
Physical properties		
Particle size distribution (coSand, meSand, fiSand, Silt, Clay)	%	Hydrometer/ Pipette method
Hydrological Properties		
Resistivity	Ohms	NS
Electrical Conductivity (EC)	mS/m	NS
Water Retentivity (-33kPa, -80kPa, -500kPa, -1500kPa)	%	NS
Relative soil saturation (AWR)		NS
MR	KPa	NS

NS- Not Specific
DTPA- Diethylenediamine Penta acetic acid
coSand- Coarse Sand
meSand- Medium Sand
fiSand- Fine Sand

Table 4-4: Chemical, and physical properties and the accompanying units of measurements as recorded in the NLEIP Report (Van Tot *et al*., 2018).

Soil Attribute	Units of measurements	Methods of measurement
Chemical Properties		
Basic Cations (Ca, Mg, Na, K)	mg/kg and cmol (+)/kg	1.0 N Ammonium acetate
Micronutrients (Zn, Mn, Cu, Co, B)	mg/kg	Di-ammonium EDTA
pH	H_20	1:2.5 water suspension
US.KCl	cmol(+)/kg	NS
Acid saturation	%	NS
Sulphate (SO4)	mg/l	NS
Sulphur (S)	mg/kg	NS
Phosphorus (S)	mg/kg	Bray 1
Cation Ratios	%	NS
Cation Exchange Capacity (CEC)	cmol(+)/kg	NS
Organic Carbon (OC)	%	Walkley-Black
Exchange Sodium Percentage **(ESP)**	%	NS
Physical Properties		
Electrical conductivity	mS/m	NS
Texture (Clay, Silt, Sand)	%	Pipette method
Dispersion (Clay, Silt, Sand)	%	Pipette method
Dispersibility Ratio (DR)		NS
Clay Flocculation Index (CFI)		NS
Clay Dispersion Ratio (CDR		NS
Hydrological Properties		
Sodium Adsorption Ratio (SAR)		NS

EDTA- Ethylenediamine tetra acetic acid
NS- Not Specified

Table 4-5: Chemical, physical, and hydrological properties and the accompanying units and methods of measurement as recorded in the HOSASH database (Le Roux *et al*., 2015).

Soil Attribute	Units of measurement	Methods of measurement
Chemical Properties		
Extractable Cations (Ca, Mg, K, Na)	cmol (+)/ kg soil	NS
Cation Exchange Capacity (CEC)	cmol (+)/ kg soil	NS
pH	H_2O/KCl	NS
Organic Carbon	%	NS
Physical Properties		
Bulk Density	g.cm^3	NS
Texture (coSa, mesa, fiSa, vfSa, Si, fiSi, Cl)	%	NS
Hydrological Properties		
Hydraulic conductivity (0mm, 30mm, 80mm, 150mm)	mm.h^{-1}	NS
Water retention (0mm, 38mm, 50mm, 100mm, 200mm, 400mm, 600mm, 800mm, 1000mm, 8000mm, 10000mm, 15000mm)	Mm	NS

Table 4-6: Physical properties and hydrological (Laboratory and IN-SITU measured properties) and the accompanying units of measurements as recorded in the Hydrological Processes Reports (Lorentz *et al*., 2001).

Soil Attribute	**Unit of measurement**
Laboratory measured properties	
Texture	
Density	kg.m^3
Porosity	m/m
Draining Upper Limit (DUL)	m/m
Wilting Point	m/m
Saturated Hydraulic Conductivity (Ks)	mm.h^{-1}
IN-SITU measurements	
Saturated Hydraulic Conductivity (Ks)	mm.h^{-1}
Hydraulic conductivity K (60)	mm.h^{-1}
Hydraulic conductivity K (150)	mm.h^{-1}

Table 4-7: Chemical, and hydrological and the accompanying units of measurements as recorded in the Data of hydraulic database (Van Tol, 2021).

Soil Attribute	**Units of measurements**
Chemical Properties	
Organic Carbon	%
Cation Exchange Capacity	cmol (+)/ kg soil
Physical Properties	
Bulk density	mg.m^{-3}
Hydrological Properties	
Hydraulic conductivity	mm.h^{-1}
Macropore conductivity	mm.h^{-1}
Draining Upper Limit (DUL)	mm.mm^{-1}
Liquid limit (LL)	mm.mm^{-1}

4.3.10 Content of the NWU database

4.3.10.1 Summary of data records from data sources

The final product of the NWU soil database was composed from three academic transcripts, 18 reports, three Research reports and two databases. These resulted in a total of 539 soil profiles and 1518 soil points. A summary of this data is displayed in table 4.8 below. Furthermore, Table 4.8 provides information about the reference of each data source, the area where soil points were collected, the total number of soil profiles and horizons acquired from each data source.

4.3.10.2 Summary of individual soil properties recorded

The charts below display the number of soil profile ID Information, morphological and physical properties (Figure 4.1); chemical properties (Figure 4.2); hydrological and geological properties (Figure 4.3) collected from different data sources and recorded in the NWU database.

Figure 4.1 cluster bar indicating that the 'name of the provider' was received from all the data sources and the least information that was provided for profile id., morphological and physical properties was the colour of dry soil with only 29 entries recorded. Coarse sand, medium sand and fine sand fractions remained constant, as the same number of data entries were recorded for these properties. Figure 4.2 shows a cluster bar indicating that the highest data entry for the chemical properties was CEC with 421 data entries and the least chemical property was cobalt (Co) as there were no data entries for both the di-ammonium EDTA and DTPA methods. The cations and micronutrients display similar amounts of data entry in each method of measurement used and 60% of the chemical properties had less than 50 data entries. The highest number of recorded hydrological properties was the saturated hydraulic conductivity, with 328 data entries and the least number of records was water retention measured at different atmospheric pressures with only one data entry recorded at each atmospheric pressure (Figure 4.3). Hydraulic conductivity at 30 mm/h^{-1} and 80 mm/h-1 have the same number of data records. Parent material and/ or Supergroup show 333 data entries while XRD mineralogy shows only 118 data entries. Overall, the morphological and physical properties had the highest number of data entries reaching over a thousand records while the chemical, hydrological and geological data entries did not reach 500 records.

Table 4-8: Data sources populated in the NWU soil database, with the accompanying reference, area, number of profiles and horizons recorded for each data source.

Source data	**Reference**	**Area**	**Total number of soil profiles**	**Total number of soil horizons**
ARC-ISCW	ARC-ISCW, 2021	Gauteng, Northwest,	19	60
HOSASH	Le Roux *et al*., 2015	Limpopo, Mpumalanga, Gauteng, Free state, Eastern Cape	51	174
PhD Thesis	Manyevere, 2014	Eastern Cape	25	76
NLEIP Report	Van Tol *et al*., 2018	Eastern Cape	49	109
DSA	Various Report	Limpopo, Mpumalanga, Gauteng, Free State, KwaZulu-Natal, Northern Cape	153	421
Hydrological Modelling data	Malan, 2016	Western Cape	38	109
Goukou Data	Smit, 2021	Western Cape	24	57
Data of hydrological properties	Van Tol, 2021	Limpopo, Mpumalanga, Gauteng, Free State, KwaZulu-Natal, Eastern Cape, Western Cape	87	191
Hydrological Processes Report	Lorentz *et al*., 2001.		93	321
TOTAL			539	1518

NLEIP- Ntabelanga and Lalini Ecological Infrastructure Project
DSA- Digital Soils Africa

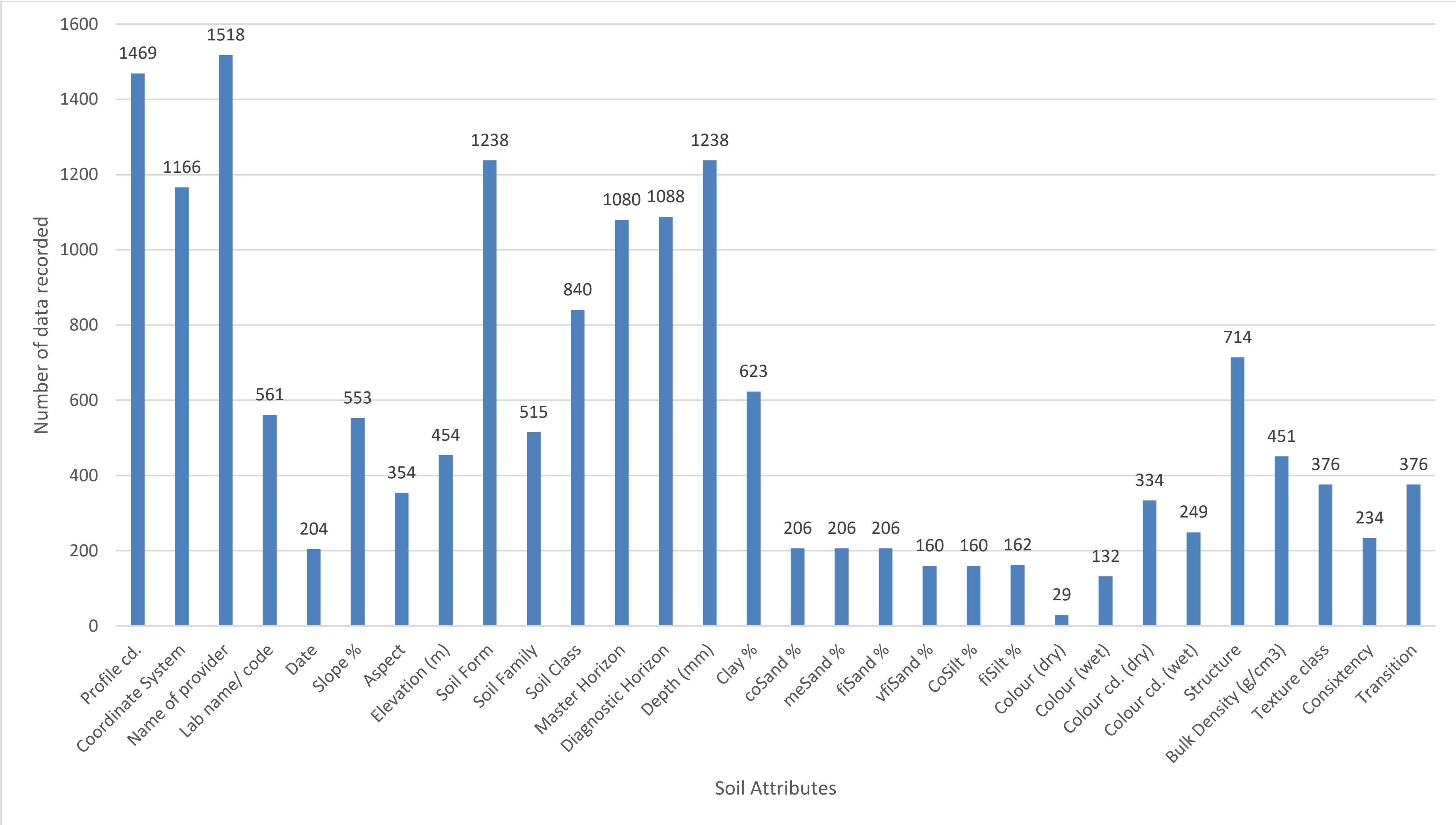


Figure 4-1: The number of profile id., morphological and physical properties recorded in the NWU soil database.

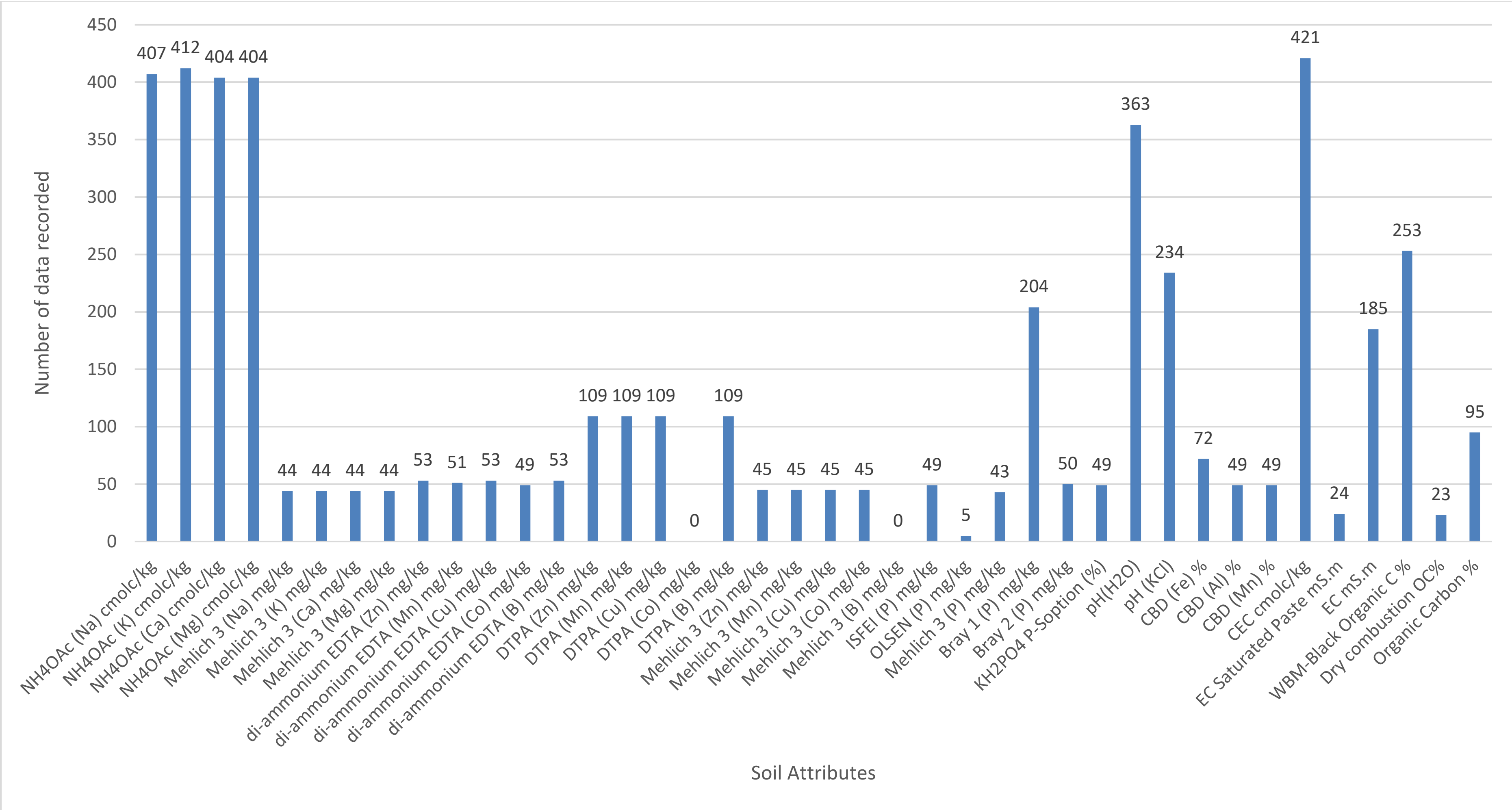


Figure 4-2: The number of chemical properties recorded in the NWU soil database.

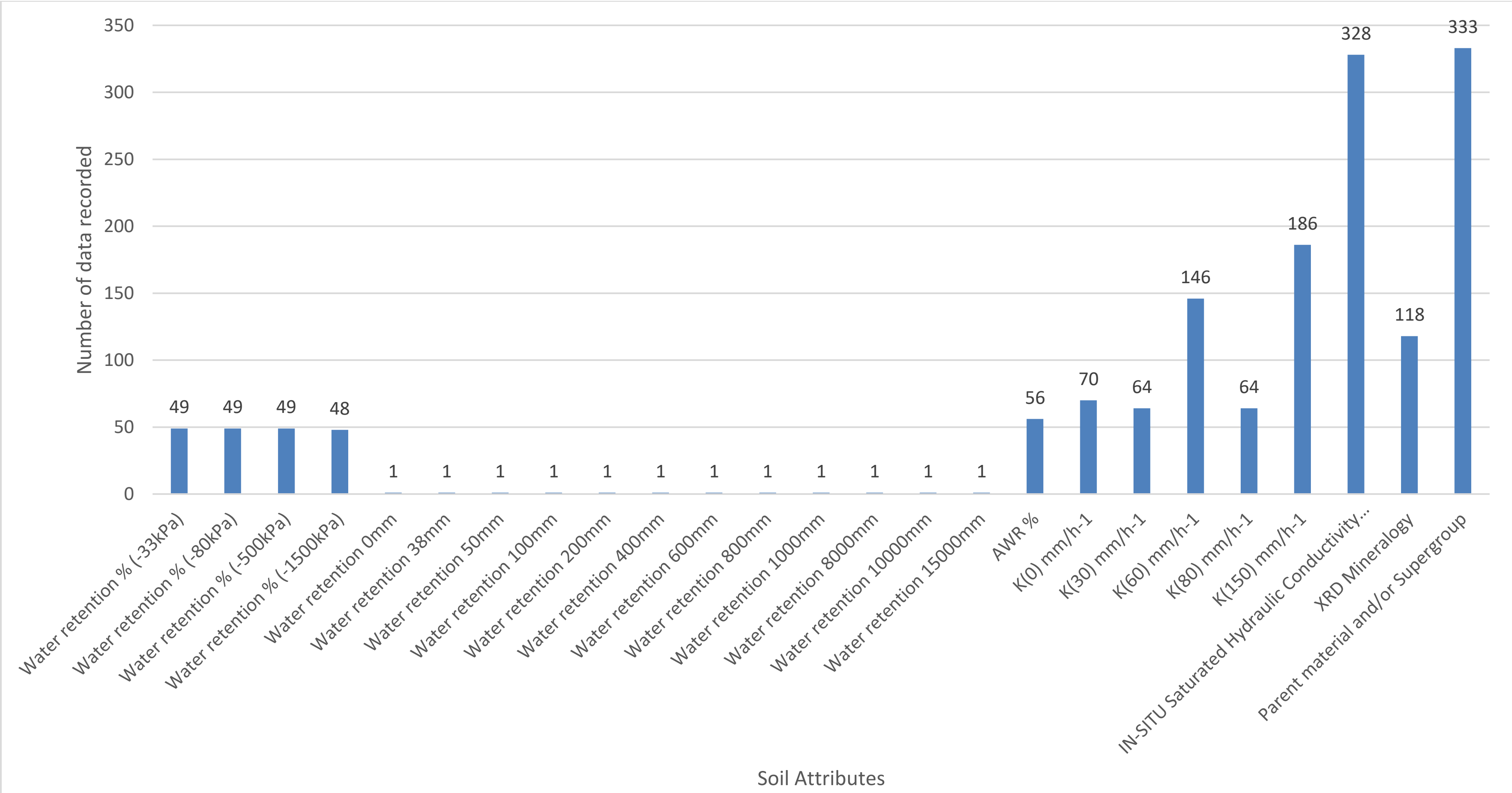


Figure 4-3: The number of hydrological and geological properties recorded in the NWU soil databa

4.4 Conclusion

Soil data from a total of 25 data sources resulting in a total of 539 soil profiles and 1518 soil horizons were incorporated in the NWU soil database. Prior to this, each data source was analysed for quality to ensure that the minimum soil information was provided for inclusion in the NWU database. The criteria used to add soil data into the NWU soil database include the presence of the name of the provider, at least one soil attribute and a unit and/or method of measurement. Standardisation of the data was performed to record the data using common units of measurements for data harmonisation. This process included conversion of geographical co-ordinates to DD, conversion of ammonium acetate measured cations from mg/kg to cmol (+)/kg, calculation of geographical co-ordinates using ARC-GIS tools and basic conversion of units for depth and bulk density to the standard units of measurements used for NWU soil database. The structure and content of each data source were described. Standardisation process that was done and the challenges encountered for each data source during data population were discussed. The number of each soil attribute for each soil property was investigated and presented in the form of cluster bar which indicated that the highest number of recorded data was collected from morphological and physical soil data.

4.5 References

ARC-ISCW. 2021. Extraction of the land type soil profile databases. ARC-ISCW, Pretoria.

Dobos, E., Bialkó, Micheli, E. & Kobza, J. 2010. Legacy Soil Data Harmonization and Database Development. Digital Soil Mapping. In J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink, S. Kienast-Brown (Eds.), *Digital Soil Mapping. Progress in Soil Science* (pp 309-323). Springer, Dordrecht.

Hortensius, D. & Norcliff, S. 1991. International standardization of soil quality measurement procedures for the purpose of soil protection. *Soil Use and Management*, 7(3):163-166. https://doi.org/10.1111/j.1475-2743.1991.tb00869.x

Le Roux, P.A.L., Bouwer, D. & Van Zijl, G.M. 2016. Irrigation Suitability Report for a portion of land identified by Mahoebe Eiendomme, Prieska, Northern Cape Province for Henry Coetzee, DSA Report.

Le Roux, P.A.L., Hensley, M., Lorentz, S.A., Van Tol, J.J., Van Zijl, G.M., Kuenene, B.T., Bouwer, D., Freese, C.S., Tinnefeld, M. & Jacobs, C.C. 2015. Hydrology of South African Soils and Hillslopes. WRC Report No. 2021/1/15.

Lorentz, S., Goba, P. & Pretorius, J. 2001 Hydrological Processes Research: Experiments and measurements of soil hydraulic characteristics. Report of the Water Research Commission on the Project "Experimentation and Laboratory Measurement for Hydrological Process Research".

Malana, 2016. Investigating the suitability of land type information for hydrological modelling in the mountain regions of Hennesqua, South Africa. (MSc Thesis). Stellenbosch University, South Africa.

Manyevera, A. (2014). An integrated approach for the delineation of arable land and its cropping suitability under variable soil and climatic conditions in the Nkonkabe municipality, Eastern Cape, South Africa. (Thesis-PhD). University of Fort Hare, Alice.

Romero, C.C., Hoogenboo, G., Baigorria, G.A., Koo, J., Gijsma, A.J. & Wood, S. 2012. Reanalysis of a global soil database for crop and environmental modeling. *Environmental Modelling & Software*, 35:163-170.

Smit, I.E., 2021. Impacts of soil information on process based hydrological modelling in the upper Goukou catchment, South Africa. (MSc thesis). University of the Free State, Bloemfontein.

Suleaman, Y., Minasny, B., McBratney, A.B., Sarwani, M. & Sutandi, A. 2013. Hamormonizing legacy soil data for difgital soil mapping in Indonesia. *Geoderma*, 192:77-85. https://doi.org/10.1016/j.geoderma.2012.08.005

Van Tol, J.J. 2021. A database for South African soil hydrological properties. University of the Free State, South Africa.

Van Tol, J.J. & Bouwer, D. 2018. Hydropedological survey of the Der Brochen Mine Exploration project, Limpopo Province. DSA Report.

Van Tol, J.J., Van Zijl, G.M., Manyevere, A., Du Plesis, C.J. & Kanuka, G. 2018. Characterisation and mapping of the erodibility of soils in selected land types in support of the Ntabalanga and Laleni ecological infrastructure project (NLEIP*)*. University of Fort Hare, Alice.

Van Tol, J.J. & Bouwer, D. 2019. Hydropedological survey of the Vopak site, Port of Richards Bay, KwaZulu-Natal. DSA Report.

Van Tol, J.J. & Le Roux, P.A.L. 2015. Hydropedological Survey Report-ABB, Nigel. DSA Report.

Van Tol, J.J. & Le Roux, P.A.L. 2019. Irrigation suitability report for the farm Harrisdale portion 223-Kilmorey, near Kimberley, Northern Cape Province for Turn180. DSA Report.

Van Tol, J.J. & Marx, J.D. 2020. Irrigation suitability of Bucklands portion 15 & 16, Douglas, Northern Cape. DSA Report.

Van Tol, J.J. & Marx, J.D. 2020. Irrigation suitability of Selene, Douglas, Northern Cape. DSA Report.

Van Tol, J.J. & Marx, J.D. 2020. WAD 510 (Klooffontein) Irrigation suitability soil survey. DSA Report.

Van Tol, J.J. & Marx, J.D. 2021. Report on the Irrigation suitability of Kameelhoek, Postmasburg, Northern Cape. DSA Report.

Van Tol, J.J. & Marx, J.D. 2021. Irrigation suitability of Willowdene, Barkly-West. DSA Report.

Van Tol, J.J. 2012. Database of hydraulic properties. University of Free State: Bloemfontein.

Van Zilj, G.M. 2015. Irrigation Suitability Report on Kheis 291 Groblershoop Northern Cape Province for Wikus Snyman. DSA Report.

Van Zijl, G.M., Bouwer, B. & Du Plessis, C.J. 2017. Irrigation Suitability Report for the Fieldsview Irrigation Project, Riverton, Northern Cape Province for Milnex. DSA Report.

Van Zijl, G.M. & Du Plessis, C.J. 2015. Irrigation Suitability Report on Fam 484 near Kolomela Mine Postmasburg Northern Cape Province for Umhlaba Consulting. DSA Report.

Van Zijl, G.M. & Du Plessis, C.J. 2016. Irrigation Suitability Report for the farm Backhouse, Douglas, Northern Cape Province for Wim van Bergen. DSA Report.

Van Zijl, G.M. & Du Plessis, C.J. 2017. Irrigation Suitability Report for the Botha Farm, Prieska, Northern Cape Province for Jan-Philip Botha. DSA Report.

Van Zijl, G.M. & Du Plessis, C.J. 2017. Irrigation suitability Report for the farm Witdam, near Kimberley, North Cape Province for van Deventer Familie Trust. DSA Report.

Van Zijl, G.M. & Du Plessis, C.J. 2017. Irrigation Suitability Report to cultivate Pecan nuts on a portion of Spitskop farm, Prieska, Northern Cape Province for Johan Gouws. DSA Report.

Van Zijl, G.M. & Du Plessis, C.J. 2017. Irrigation Suitability Report to cultivate Pecan nuts on a portion of Uitdraai farm, Prieska, Northern Cape Province for Orffer Muller. DSA Report.

Suleaman, Y., Minasny, B., McBratney, A.B., Sarwani, M. & Sutandi, A. 2013. Hamormonizing legacy soil data for difgital soil mapping in Indonesia. *Geoderma*, 192:77-85. https://doi.org/10.1016/j.geoderma.2012.08.005

CHAPTER 5 DATA QUALITY CONTROL

5.1 Introduction

Statistical analysis such as outlier detection and correlation tests are used to investigate the quality of data. Over the years, outlier detection has been used to find and, if necessary, delete abnormal occurrences from data. An outlying observation, also known as an outlier, is one that appears to differ significantly from other datasets in which it occurs (Hodge & Austin, 2004:86). Although statistical analysis is used to detect outliers, these points may not have erroneous anomalies capable of affecting the quality of the data but can instead transmit useful information about the data's dynamics (Filzmoser & Gregorich, 2010:1051). After the population of data into the NWU soil database, this data needed to be quality controlled to ensure accuracy and data quality. The objective of this chapter is to perform quality control to decide which data is accurate enough to be recorded in the NWU soil database.

5.2 Materials and Methods

Quantitative soil properties, including bulk density, soil texture, cations, micronutrients, Phosphorus status, Phosphorus-sorption, pH, metal ions, electrical conductivity, Organic Carbon, water retention, relative soil saturation, hydraulic conductivity, and saturate hydraulic conductivity, were quality controlled. Basic quality control was performed through observation of the soil database analytical properties based on the criteria set by Leenaars (2013) and Batjes (2008).

5.2.1 Basic quality control

Quantitative soil properties have been quality controlled in the NWU soil database and this process was characterised by three stages. The first stage of quality control involved removal of all the values detected as outliers from the basic quality control criteria. The following criteria was used as adopted from the WISE 3 (Batjes, 2008) and African Soil Profiles Database version 1.1 (Leenaars, 2013):

1. Identify and exclude unknown geographical co-ordinates or soil co-ordinates not falling within the borders of South Africa.
2. Identify values for the sum of sand, silt and clay fractions less than 90% and exceeding 110%.
3. Identifying bulk density values outside the range of 0.2-2.7g/cm^3
4. Identifying exchangeable calcium outside the range of 0-100 cmolc/kg, magnesium outside the range of 0-50 cmolc/kg, sodium outside the range of 0-100 cmolc/kg and potassium outside the range of 0- 20 cmol/kg.
5. Identify Sum of Exchangeable bases values exceeding 150 cmol/kg
6. Identifying P-status values outside the range of 0-1000 mg/kg.

7. Identifying pH (H_2O) outside ranges 2 - 12 and pH (KCl) exceeding pH (H_2O) values.
8. Identifying soil properties values expressed with a percentage value (e.g., clay, silt, sand, Organic Carbon, P-sorption, Fe, Al, Mn, water retention, AWR), exceeding 100%.
9. Identifying CEC values exceeding outside range 1-120 cmolc/kg
10. Identifying EC values <0 mS/m, EC values exceeding 3000 mS/m while pH (H_2O) <7.5 or pH (KCl) < 6.

5.2.2 Outlier Detection

The second stage of quality control involved the detection of outliers with boxplots using the SPSS software. The boxplot presented two groups of outliers- ***extreme outliers*** depicted by a star and plotted beyond the upper boundary of the boxplot, and ***mild outliers*** depicted by a circle, plotted between upper cut off and upper boundary of the boxplot (Li *et al*., 2020: 564). All these values were flagged (Figure 5.2-5.6, and then investigated for potential errors in the following manner:

1. Revising values- this process involved revisiting the data source to ensure that the values were copied correctly.
2. Analysis of surrounding values-a value detected as an outlier was analysed and compared to the surrounding values to investigate if there were any similarities, differences, or trends.
3. Comparison of values with those observed in the literature.
4. Analysis of geological properties for chemical values detected as outliers - a value detected as an outlier and is different from the surrounding values would go through further analysis where the geology would be investigated as underlying geology has the potential to alter the chemistry (chemical properties) of soil (Djodjic, *et al*., 2021)
5. Analysis of terrain position- a value detected as an outlier and is similar to the surrounding values and not affected by the underlying geology would be investigated to analyse if the terrain position has any impact on the value, as the terrain slope position may have an influence on soil chemical properties (Assefa, *et al*., 2020:14).

5.3 Results and Discussion

5.3.1 Basic quality control

5.3.1.1 Geographical co-ordinates

Geographical co-ordinate analysis was accomplished through the development of a map in ArcMap (Figure 5.1). From the created map it was observed that all the data points excluding a few soil points from the HOSASH (Le Roux *et al*., 2015), and PhD Thesis from (Manyevere, 2014) datasets database were plotted within the borders of South Africa in specific provinces where data sampling was carried out. Since all the sample points were plotted in GIS, therefore, all the soil points with co-ordinates have a known geolocation. However, five geographical co-ordinates from the HOSASH (Le Roux *et al*., 2015) database and two geographical co-ordinates from Manyevere (2014) did not plot within the borders of South Africa (Figure 5.1). Therefore, the sources were revisited to check if the coordinates were recorded properly and because this was the case, the co-ordinates were omitted from the database.

5.3.1.2 Sum of coarse fragments (≤110%)

A total of 13 soil points had values with the sum of particle size exceeding 110% (Annexure C). These values were removed from the NWU soil database.

5.3.1.3 CEC values (≤ 120 cmol/kg)

A total of seven soil point values were flagged as the points expressed values of CEC exceeding 120 cmol/kg (Annexure C).

5.3.1.4 Soil property values adhering to criteria

Bulk density recorded values within the range 0.2 – 2.7 g.cm^{3}and EC also recorded values within the expected range as stipulated in the criteria. All the cations recorded values within the given criteria for each cation, calcium (0 - 100 cmolc/kg), magnesium (0 - 50 cmolc/kg), sodium (0 - 100 cmolc/kg) and potassium (0 - 20cmol/kg). Furthermore, the total sum of all the values recorded for each cation did not exceed 150 cmolc/kg, meaning that values for exchangeable bases were retained in the NWU soil database. P -status values also recorded within the expected range (0 - 1000 mg/kg). All the pH (H_2O) and pH (KCl) were within the range of 2 - 12 as set in the criteria. Organic Carbon, P-sorption, Fe, Al, Mn, water retention, AWR values recorded in percentages did not exceed 100% as stipulated in the criteria. Therefore, the values for these properties were retained in the NWU soil database.

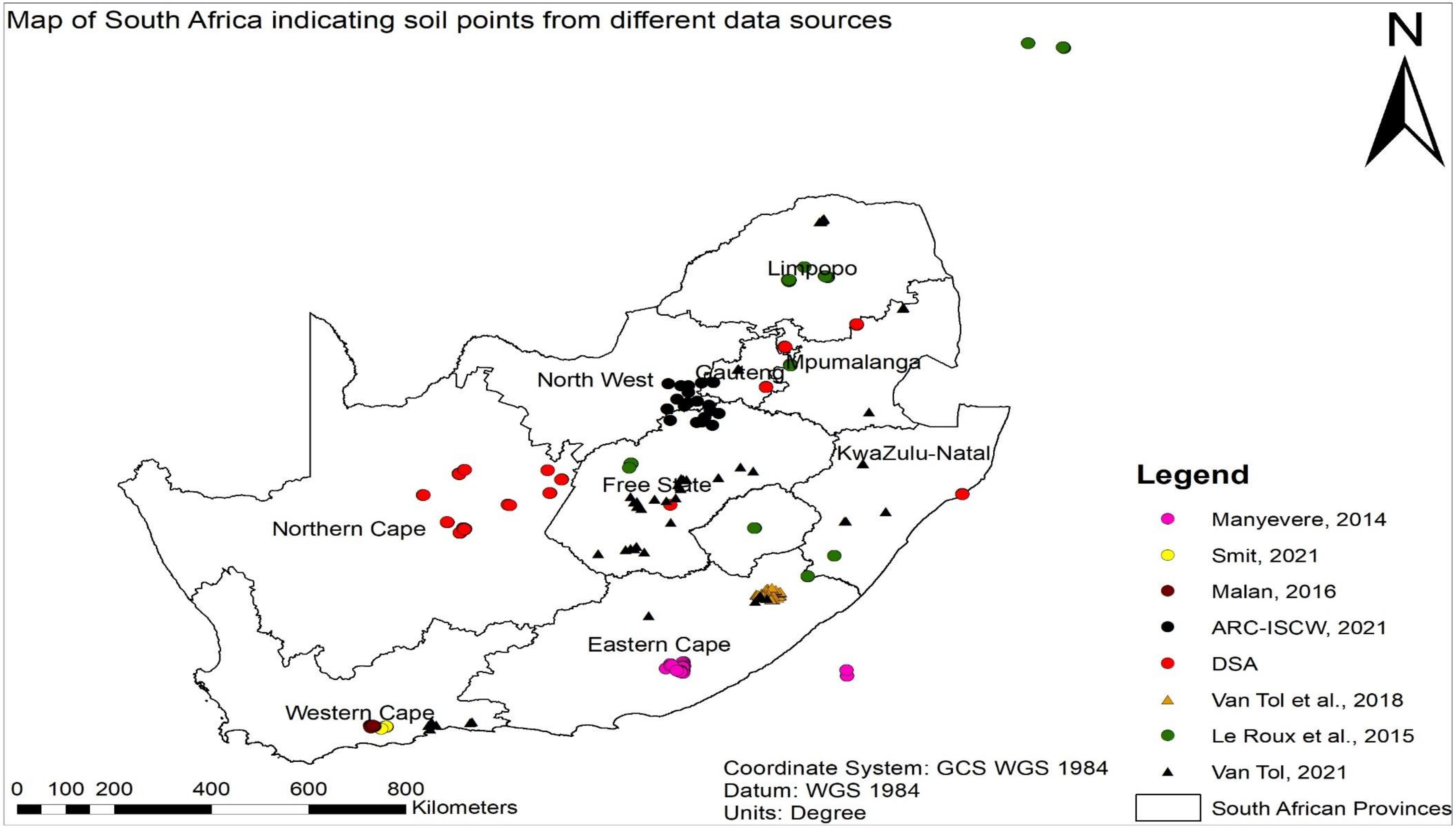


Figure 5-1: Map of South Africa indicating the location of soil points from different data sources.

5.3.2 Outlier Detection

Boxplots were created for all the quantitative soil properties (Figure 5.2 to Figure 5.6). The boxplots presented the outliers as case numbers (case number is the row number where the value was recorded in the dataset), while outlier values recorded for the boxplots are displayed in Annexure C. No outliers were detected for bulk density values.

5.3.2.1 Soil Texture

One mild outlier was detected for clay content, this point was detected as an outlier as it was the highest value of clay 68.70% in the database with an accompanying CEC value of 34.10 cmol/kg and this value was average (not high or low) as compared to the other CEC values. There was no geological and terrain data to further analyse this value was detected based on the surrounding values. Furthermore, Bonheim soil forms are expected to have clay % above 35 (MaVicar & Perfect, 1971: 194). Therefore, this value was not excluded from the database as it does not affect the quality of the database in anyway.

Eight mild outliers and one extreme outlier were detected for coSand %. The mild outliers were high as compared to the surrounding data, thus detected as outliers and the extreme outlier was the highest value of 53% recorded for coSand %. As stipulated in the soil database the texture class of the specific soil horizon is sandy and soil with >53% sand and 2.20% clay is classified as sandy soil according to the soil texture chart (Fertilizer Society of South Africa, 2007:16). Furthermore, thick alluvium sands can occur in environments underlain by Dwyka tillite (Gomo & Masemola, 2016:3). Therefore, none of these values were excluded from the NWU Soil database.

Six mild outliers and five extreme outliers were detected for meSand%. Values detected as mild outliers are average as compared to the surrounding values and those detected as extreme outliers are higher as compared to the surroundings values. The accompanying sand fractions display low values of sand %, and similar to the coSand%, the soil horizon is classified as a sandy texture class and underlain by Dwyka tillite.

Two mild outliers were detected for fiSand %. These values were slightly higher than the surrounding values but were not too different from the surrounding values to be flagged as outliers. Therefore, these values were not excluded from the NWU soil database. Ten mild outliers were detected for vfiSand %. Similarly, the values were slightly higher than the surrounding values and in these cases these values were high due to the low values recorded for the other soil fractions for the same soil horizon. All these horizons are characterised by sandy texture class. Therefore, these values were not excluded from the database.

Two mild outliers and six extreme outliers detected for coSilt%. Case no. 92 of detected mild outliers is slightly higher than the surrounding values and the other soil texture fractions, however, the total Clay, Sand, and Silt % sums up to 100%. Therefore, this value was not excluded from the database. Case no. 231 was detected as a mild outlier, however, the value was already removed based on the criteria set for basic quality control analysis, as the %sum of soil texture fractions was >110%. Therefore, all the mild outliers detected for case no. 231 including vfiSand%, coSilt% and fiSilt % were excluded from the soil database as they affected the sum of soil fractions. The extreme outlier detected from case no. 73 recorded a value of 115%, this value was already removed based on the criteria set for basic quality control. Cases no. 206, 222, 223, and 227 were also detected as an outlier in both the basic quality control and outlier detection. The %sum of the soil particle size was >110%, thus soil texture values for this case no. were omitted form the database.

Six mild outliers and three extreme outliers were detected for fiSilt%. Case no. 112 was already detected as an outlier from the basic quality control as the %sum of soil texture was >110%. Cases no. 117, 149,172, recorded similar values for fiSilt%, these values and case no. 237 were slightly higher as compared to the surrounding values. However, the % sum in this case number was <102%, therefore these values were not excluded from the database. Case no. 227 was already removed from the database as the %sum of soil texture in this case number was >110%.

5.3.2.2 Soil chemical properties and hydrological properties.

During analysis of chemical properties detected as potential outliers, it was observed that the extreme outliers for di-ammonium EDTA (Mn) were high as compared to the surrounding values and the other di-ammonium EDTA measured elements. The soil point did not contain terrain information for further analysis and the geology information was the same as that of the surrounding values. The recorded high values for di-ammonium EDTA (Mn) were substantiated by fact that high amounts of Mn averaged at 650 mg/kg are expected in soils as Mn is one of the ten most abundant elements in the upper continental crust (Gilkes & McKenzie, 1988:23). The three mild outliers detected for di-ammonium EDTA (Co) were represented by each soil horizon within the same soil profile. Therefore, this anomaly from other surrounding values may be due to the spatial variability of soil. The extreme outliers detected for EC, may be due to the terrain positions as some of the points were collected at the lowest point of a depression, possibly resulting in salt accumulation in this area (Van Tol & Marx, 2020).

Other chemical properties-NH_4OAc (Na, K, Ca, Mg), Mehlich 3 (Na, K, Ca, Mg), di-ammonium EDTA (Zn, Co, B), DTPA (Zn, Mu, Cu, Co, B), Mehlich 3 (Zn, Mu, Cu, Co, B), P-status, CBD (Fe, Mn, Al %) and CEC and hydrological properties (water retention, AWR %, hydraulic conductivity, and saturated hydraulic conductivity), were detected as outliers. However, these values were not considered as true outliers because:

1. The values were one unit (sometimes less) lower as compared to the surrounding values. Therefore, outlier values were slightly higher or lower than the surrounding values, outlier values were not extremely different from values not detected as outliers.
2. During the analysis of the geology and terrain of the values detected as outliers, it was observed that the geology and terrain of these values were the same as that of the values not detected as outliers. Therefore, if the effect of the outlier detected was caused by terrain and/or geology this would have affected the surrounding values similarly.
3. Values with no terrain and geology information were analysed by plotting the outliers on a map to investigate the origin of the soil point. Similarly, some of the values detected as outliers were in close proximity (same position and underlying material) to the values not detected as outliers, implying that the geology and terrain had no impact on the outlier values.

Due to the mentioned reasons, detected outliers cannot be considered as true outliers as there is not enough substantiation of this output. As stated by Filzmoser & Gregorich (2010: 1051), points within a dataset that are detected as outliers may not be erroneous anomalies with the potential to affect the quality of the dataset. All the values were not removed from the database because in each case it was deemed that the flagged values were likely to occur based on the five-step verification process.

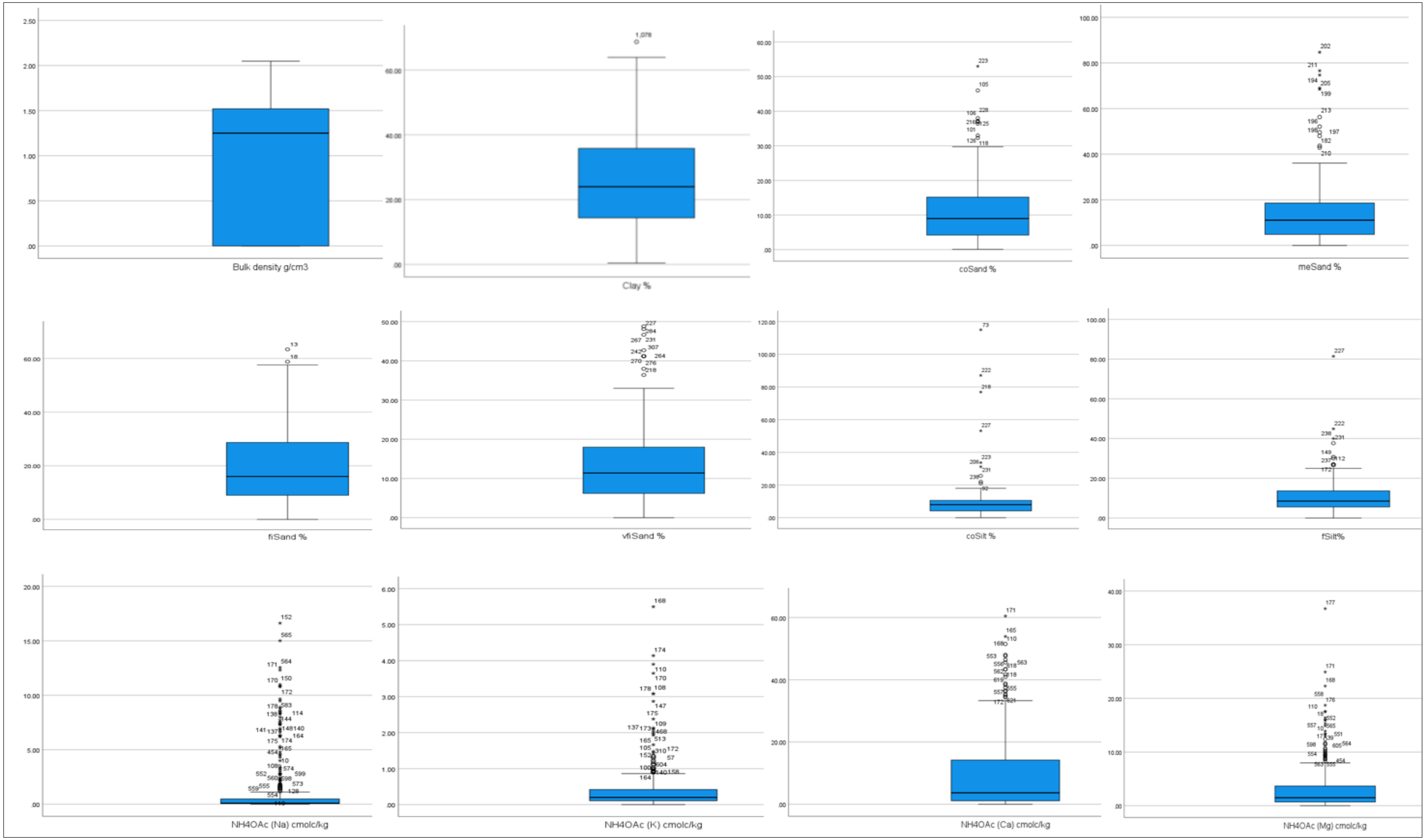


Figure 5-2: Boxplots displaying the extreme and mild outliers detected for bulk density, soil texture classes and NH_4OAc (Na, K, Ca, Mg). The numbers given to outliers refer to the row number of the data point.

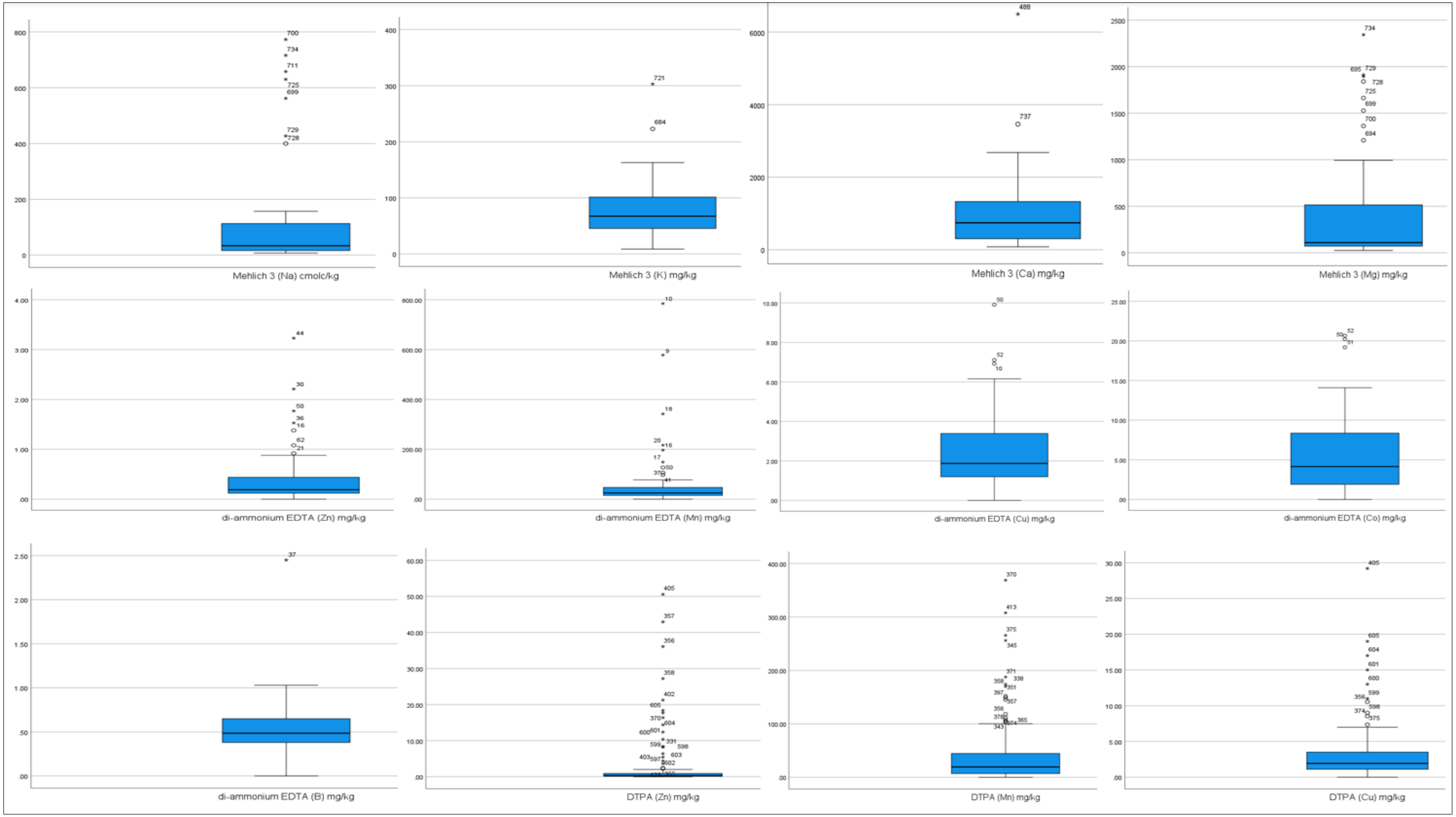


Figure 5-3: Boxplots displaying the extreme and mild outliers detected for bulk density, Mehlich 3 (Na, K, Ca, Mg), di-ammonium EDTA (Zn, Mn, Cu, Co, B) and DTPA (Zn, Mn, Cu). The numbers given to outliers refer to the row number of the data point.

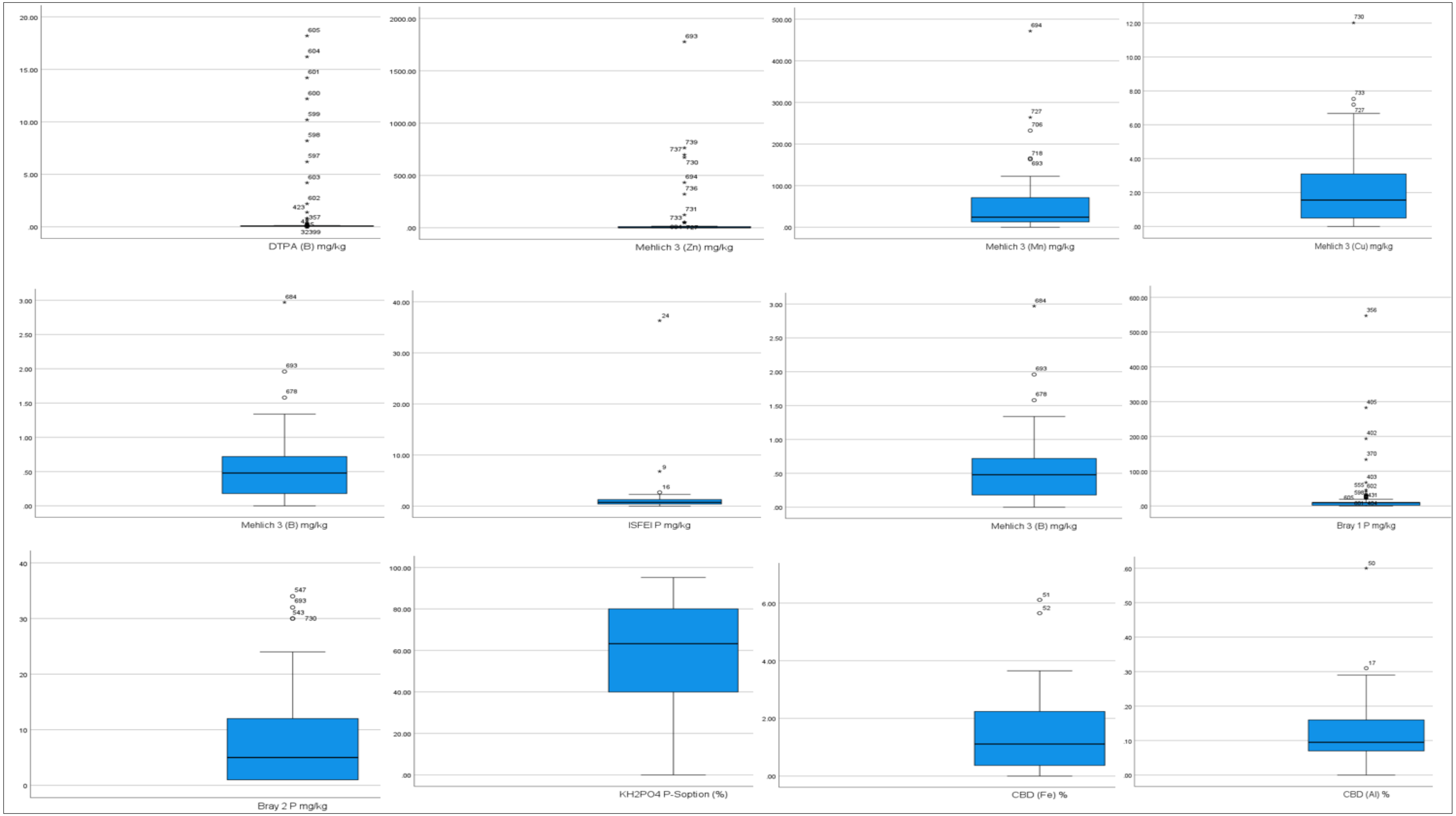


Figure 5-4: **Boxplots displaying the extreme and mild outliers detected for and DTPA (B), Mehlich (Zn, Mn, Cu,B,), OLSEN P, Mehlich P, Bray 1 P, Bray 2 P, KH_4PO_4 P-Sorption, CBD (Fe, Al) %. The numbers given to outliers refer to the row number of the data point.**

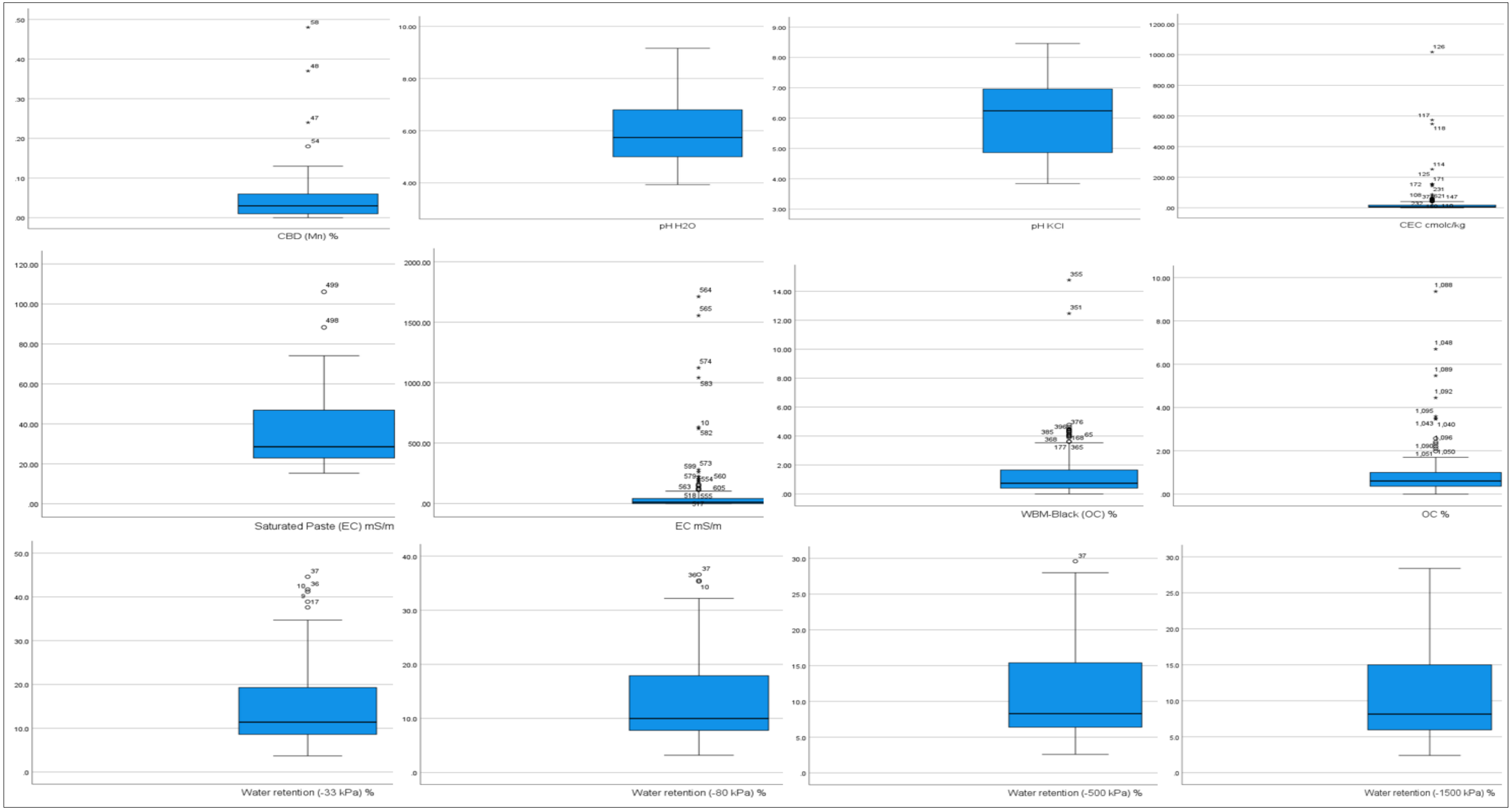


Figure 5-5: Boxplots displaying the extreme and mild outliers detected for CBD (Mn) %, pH (H_20), pH (KCl), CEC, Saturated Paste (EC), EC, WBM-Black (OC)%, OC%, Water retention (-33, -80, -500, -1500 kPa). The numbers given to outliers refer to the row number of the data point.

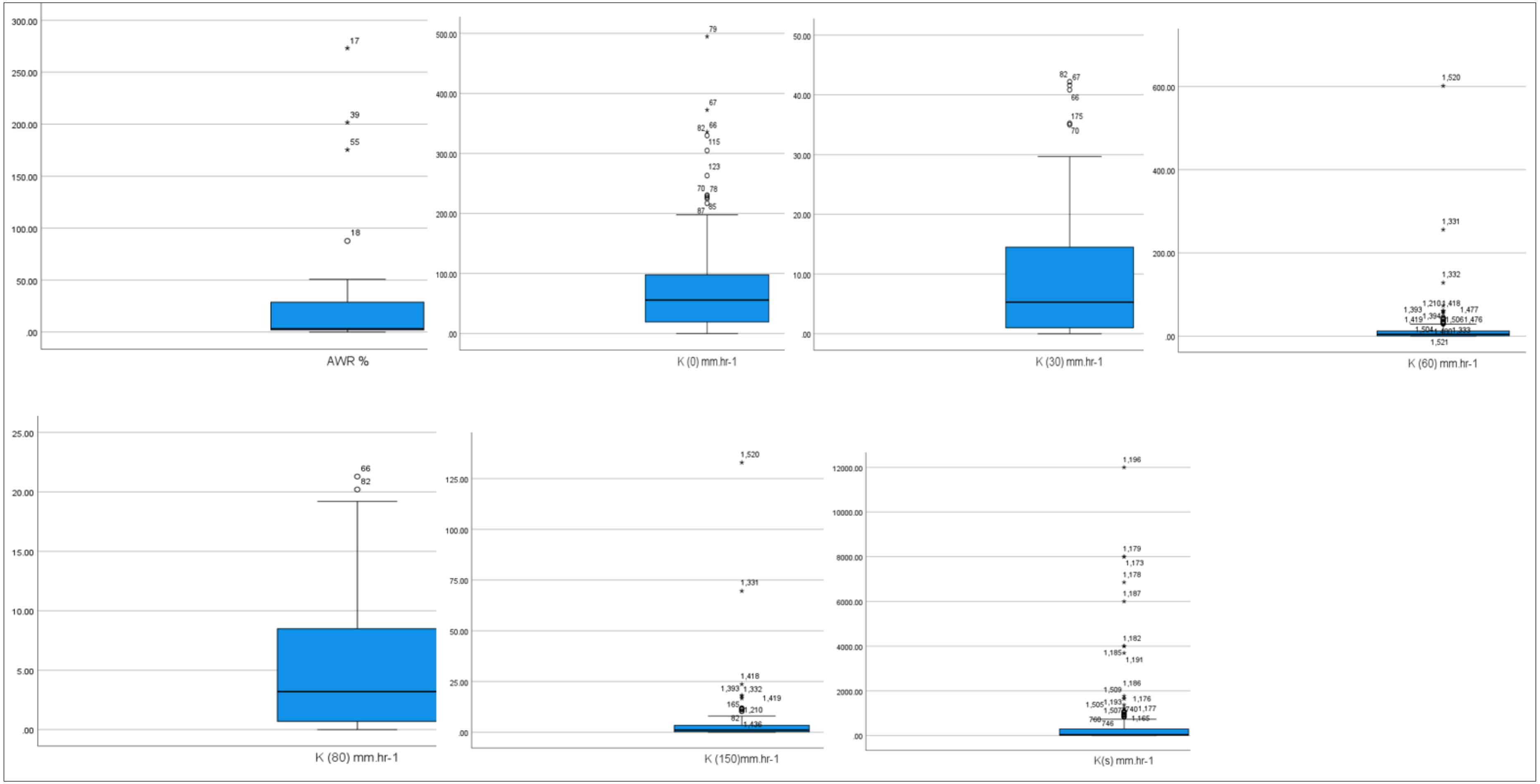


Figure 5-6: **Boxplots displaying the extreme and mild outliers detected for AWR%, K (0), K (30), K (60), K (80), K (150) and K(s). The numbers given to outliers refer to the row number of the data point.**

5.3.3 Content of NWU soil database after quality control

Soil attributes including profile cd., name of provider, date, slope, aspect, elevation, soil form and family, soil class, master and diagnostic horizon, depth, XRD Mineralogy and parent material/Supergroup were not analysed for quality control, therefore, the number of information recorded for each attribute stayed the same.

From basic quality control analysis, a total of seven outliers were detected from the geographical co-ordinates data, 13 from the sum of particle size, one from soil properties expressed in percentages and seven from CEC. Therefore, a total of 28 soil point values were detected as outliers and omitted from the NWU soil database. Table 5.1 displayed the number of soil attributes recorded before and after quality control.

Table 5-1: Summary of the different of soil attributes recorded before and after quality control.

Soil Attributes	Before Quality control	After Quality control
Geographical co-ordinates	1166	1140
Clay %	623	611
coSand %, meSand%, fiSand%	206	194
vfiSand%, coSilt %	160	148
fiSilt%	162	150
CEC	421	414

5.4 Conclusions

After soil data population in the NWU soil database, data quality control was carried out to ensure accuracy and data quality. Quality control in the NWU soil database was carried out in two stages. The first stage was the basic quality control using the criteria as set forth in the African soil profile database and WISE 3. The second stage of quality control was based on the detection of outliers using the boxplots created in SPSS. Basic quality control resulted in the detection and exclusion of a total of 27 soil point values. Quality control performed with outlier detection did not result in any exclusion of soil point values, as in each case it was deemed that the flagged values were likely to occur based on the five steps verification process. After quality control, the data added to the NWU soil database are deemed to be of sufficient accuracy to be used with confidence. The final NWU data for still consists of 539 soil profiles with 1518 soil horizons represented.

5.5 References

Assefa, F., Elias, E., Soromessa, T. & Ayele, G.T. 2020. Effect of Changes in Land-Use Management Practices on Soil Physiochemical Properties in Kabe Watershed, Ethiopia. *Air, Soil and Water Research*, 13:1-16. https://doi.org/10.1177%2F1178622120939587

Batjes, N.H. 2008. ISRIC-WISE Harmonized Global Soil Profile Dataset (Ver. 3.1). Report 2002/02, ISRIC- World Soil Information, Wageningen. http://www.isric.org/isric/Webdocs/Docs/ISRIC_Report_2008_02.pdf

Djodjic, F., Bieroza, M. & Bergström, L. 2021. Land use, geology and soil properties control nutrient concentrations in headwater streams. *Science of the Total Environment*, 772.145108. https://doi.org/10.1016/j.scitotenv.2021.145108

Fertillizer Society of South Africa. 2007. FSSA fertilizer handbook. Lynnwood Ridge [South Africa]: Fertilizer Society of South Africa.

Filzmoser, P., Garrett, R.G. & Reimannn, C. 2005. Multivariate outlier detection in exploration geochemistry. *Computer & Geosciences*, 31: 579-587. https://doi.org/10.1016/j.cageo.2004.11.013

Gilkes, R.J. & McKenzie, R.M. 1998. Geochemistry and Mineralogy of Manganese in Soils. *Manganese in Soils and Plants*, 33: 23-35.

Gomo, M. & Masemola, E. Groundwater hydrogeochemical characteristics in rehabilitated coalmine spoils. *Journal of African Earth Sciences*, 116: 114-126. https://doi.org/10.1016/j.jafrearsci.2015.12.015

Hodge, V.J. & Austin, J. 2004. A survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22: 85-126.

Leenars, J.G.B. 2013. African Soil Profiles Database Version 1.1. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). ISRIC Report 2013/03. 1-160.

Li, B., Fan R., Tang, J., Qui, J. & Wang, Y. 2020. A Hybrid Quality Control Strategy for Consumption Data on Petroleum, Oil and Lubricants by Oil Equipment. *International Journal of Heat and Technology*, 38(2):562-568. https://doi.org/10.18280/ijht.380234

Macvicar, C.N. & Perfect, G.A. 1971. The soils of the Eastern Transvaal sugar industry. *Proceedings of The South African Technologists 'Association.* 189-195.

Van Tol, J.J. & Marx, J.D. 2020. Irrigation suitability of Selene, Douglas, Northern Cape. DSA Report.

Zgorelec, Z., Grahova, B., Percin, A., Jurkovic, V., Gandjaeva, L. & Maurovic, N. 2019. Comparison of Two Different CEC Determination Methods Regarding the Soil Properties. *Agriculturae Compectus Scientificus*, 84, 2. 151-158.

CHAPTER 6 SOIL DATA CHARACTERISATION

6.1 Introduction

Soil characterisation is an important concept as it leads to understanding of the processes occurring within soils due to the variation of chemical and physical properties (Heil & Schmidhalter, 2012:98). This further assists in the mitigation of soil degradation and the effective management of soils for sustainable agricultural productivity (Castrignanó, *et al.*, 2002: 44). The objective of this chapter is to characterise the populated and quality-controlled soil data. The purpose of the characterization was to provide information of the geographical, morphological, physical, and chemical and hydrological characterization of soil data populated in the NWU soil database.

6.2 Materials and Methods

6.2.1 Geographical characterisation of soil data

Geographical maps were created for the NWU soil database to display where the soil points were collected in South Africa. This process was carried out using ArcMap. Geographical co-ordinates recorded in the NWU soil database were entered into ArcMap and for the points to plot within the geographical space of South Africa, all the co-ordinates needed to have a common geographical system used in South Africa (WGS 1984 geographical system). Each sample was denoted by a point to show where on the map the soil point was samples or measured.

6.2.2 Morphological and physical characterisation of soil data

The morphological and physical characterisation of soil data from the NWU soil database was achieved using Soil Form, soil texture and bulk density. A graph was created in Excel displaying the number of soil forms recorded in the NWU soil database. Descriptive analysis table was created in SPSS to indicate the descriptive statistics of the physical properties (Table 6.3) together with the criteria used in Hazelton and Murphy (2007), for characterisation of the soils.

6.2.3 Chemical and hydrological characterisation of soil data

The analytical characterisation of soil data from the NWU soil database was achieved using soil attributes including NH_4OAc measured exchangeable cations, Phosphorus status, pH (H_2O), CEC, OC, and hydraulic conductivity. Descriptive analysis tables created in SPSS to indicate the descriptive statistics of the analytical properties (Table 6.3) and criteria from Hazelton and Murphy (2007), were used for soil characterisation. Table 6.1 and Table 6.2 indicates the criteria used:

Table 6-1: Criteria used for physical, chemical and hydrological soil attribute values (Hazelton & Murphy, 2007).

Ratings	Clay%, Sand%, Silt%	CEC	Na, K	Ca	Mg	P	Pb	OC	K(s)
Extremely low								<0.40	<0.5
Very low	<10	<6	0-0.1	0-2	0-0.3	<5	<1.0	0.40-0.60	0.5-10
Low	10-25	6-12	0.1-0.3	2-5	0.3-1.0	5-10	1.0-1.3	0.60-1	10-20
Moderate	25-40	12-25	0.3-0.7	5-10	1-3	10-17	1.3-1.6	1-1.80	20-60
High	40-50	25-40	0.7-2.0	10-20	3-8	17-25	1.6-1.9	1.80-3	60-120
Very high	>50	>40	>2	>20	>8	>25	>1.9	>3	>120

Table 6-2: Criteria used for pH values (Hazelton & Murphy, 2007).

Ratings	pH
Very strongly alkaline	>9.0
Strongly alkaline	9.0-8.5
Moderately alkaline	8.4-7.9
Mildly Alkaline	7.8-7.4
Neutral	7.3-6.6
Slightly acidic	6.5-6.1
Moderately acidic	6.0-5.6
Strongly acidic	5.5-5.1
Very strongly acidic	5.0-4.5

6.3 Results and discussion

6.3.1 Geographical characterisation of soil data

Soils represented in the NWU soil database were collected across the entire South African geographical space from Limpopo Province, Mpumalanga Province, Gauteng Province, North West Province, Free State Province, KwaZulu-Natal Province, Western Cape Province, Eastern Cape Province and Northern Cape Province (Figure 6.1).

6.3.2 Morphological and Physical characterisation of soil data

Out of a total of 73 soil forms of the South African Classification (Soil Classification Working Group,1991), the NWU soil database includes only 58 of the soil forms of South Africa. Most of the soil profiles in the database are characterised by the Hutton soil form, followed by Oakleaf and Katspruit (Figure 6.2). The minimum value of clay recorded is 0.4% and the maximum value recorded is 68.7%. The minimum value recorded for sand is 2.5% and the maximum value recorded is 94.4%. The values recorded range between very low to very high clay % and sand %. Therefore, the clay % and sand % values recorded in the NWU soil database encompass all classes of Hazelton and Murphy (2007). The minimum value recorded for silt is 0.8% and the maximum value recorded is 41%. The values for silt% range between very low and high silt %. Very high values of silt were not recorded in the NWU soil database.

The minimum value recorded for bulk density is 0.25 $g.cm^3$ and the maximum value recorded is 2.04 $g.cm^3$. The values range between very low and very high bulk density. Therefore, bulk density values recorded in the NWU soil database encompass all classes of Hazelton and Murphy (2007).

6.3.3 Chemical and hydrological characterisation of soil data

The minimum value recorded for Na is 0.003 cmolc/kg and maximum value recorded is 16.62 cmolc/kg. The minimum value recorded for K is 0.01 cmolc/kg, and the maximum value recorded is 5.5 cmolc/kg. The minimum value recorded for Ca is 0.03 cmolc/kg and maximum value recorded is 60.48cmolc/kg. The minimum value recorded for Mg is 0.06 cmolc/kg and the maximum value recorded is 36.72 cmolc/kg. The values range between very low and very high exchangeable cation levels. Therefore, exchangeable cation values recorded in the NWU soil database encompass all classes of Hazelton and Murphy (2007).

The minimum value of CEC is 1.61 cmolc/kg and maximum value recorded is 88 cmolc/kg. The values range between very low and very high CEC. Therefore, CEC values recorded in the NWU soil database encompass all classes of Hazelton and Murphy (2007). The minimum value recorded for pH is 3.93 and the maximum value recorded is 9.16. The values for pH range between very strongly acidic and very strongly alkaline. Therefore, pH values recorded in the NWU soil database encompass all classes of Hazelton and Murphy (2007).

The minimum value recorded for Phosphorus is 0.05 mg/kg and maximum value recorded is 547.20 mg/kg. The values recorded for Phosphorus range between very low and very high Phosphorus levels. Therefore, Phosphorus values recorded in the NWU soil database encompass all classes of Hazelton and Murphy (2007). The minimum value recorded for OC is 0.05 % and the maximum value recorded is 99.17%. The values recorded for OC range between extremely low to very high OC. Therefore, OC values recorded in the NWU soil database encompass all classes of Hazelton and Murphy (2007). The minimum value recorded for IN-SITU saturated hydraulic conductivity K(s) is 104.45 $mm.hr^{-1}$ and the maximum value recorded is 12000 $mm.hr^{-1}$. The values recorded for saturated K(s) range between high and very high K(s). No values were recorded between extremely low and moderate K(s) v

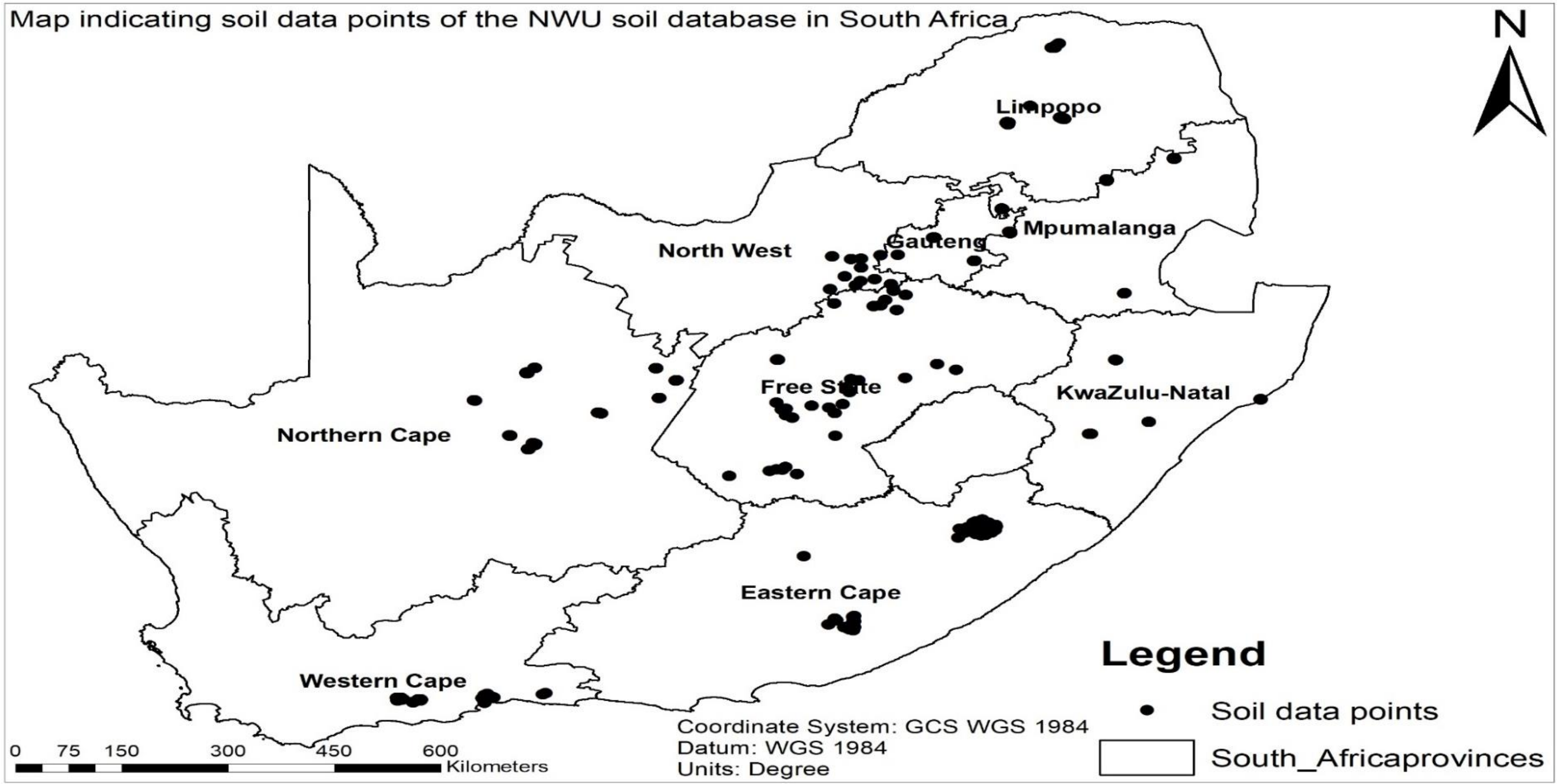


Figure 6-1: South African map indicating the soil points recorded in the NWU soil database.

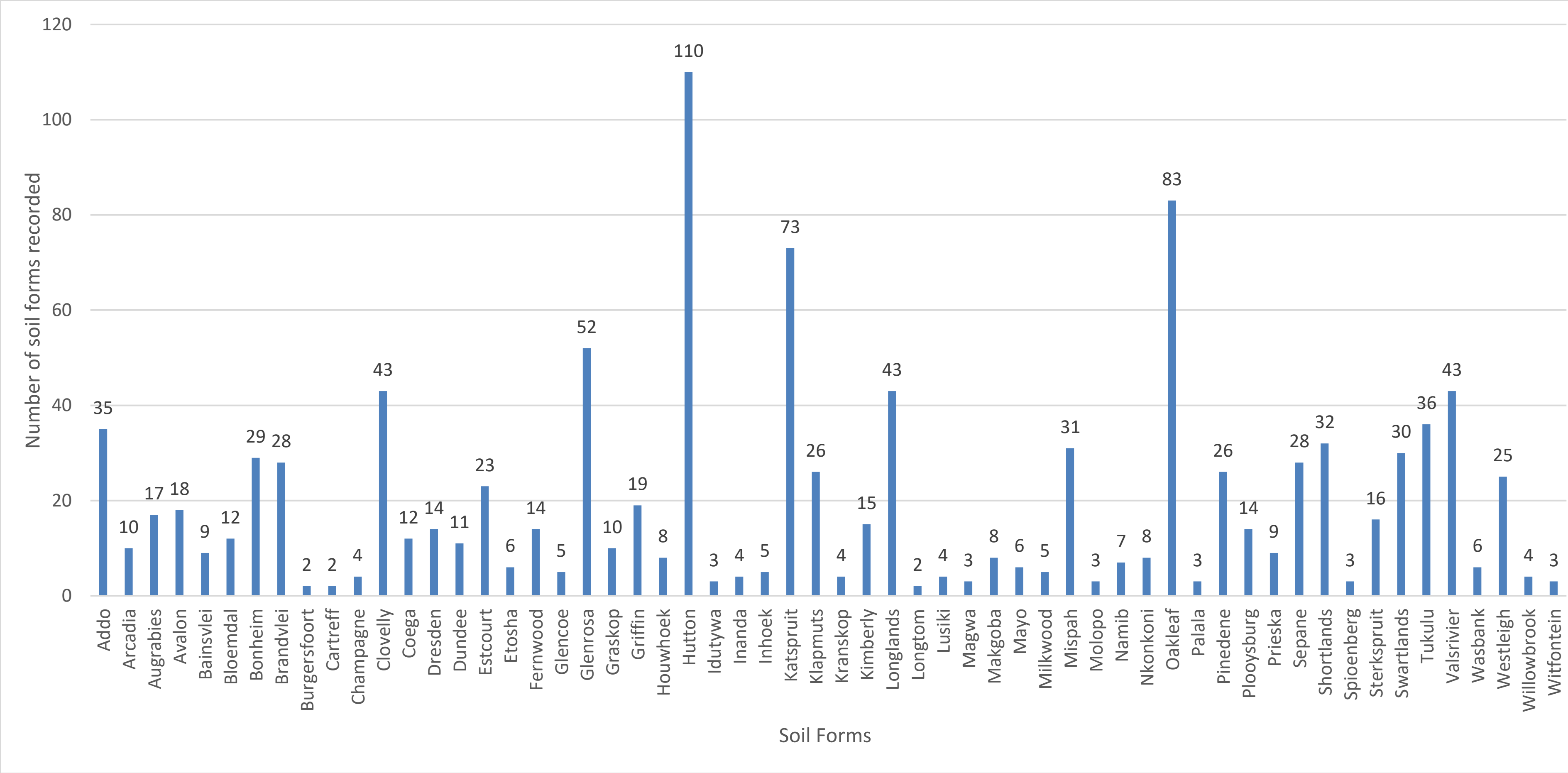


Figure 6-2: Soil forms recorded in the database and the percentage of entries per soil form.

Table 6-3: Descriptive statistic table for the physical, chemical and hydrological soil attributes.

Descriptive Statistics					
	N	Minimun	Maximum	Mean	Std. Deviation
Clay %	609	0.40	68.70	25.41	14.34
Sand %	191	2.50	94.40	56.58	21.47
Silt %	148	0.80	41.00	17.18	9.13
Bulk Density g/cm3	279	0.25	2.04	1.34	0.29
Na cmolc/kg	384	0.00	16.62	1.07	2.48
K cmolc/kg	387	0.01	5.50	0.40	0.58
Ca cmolc/kg	404	0.03	60.48	8.81	10.92
Mg cmolc/kg	401	0.06	36.72	3.03	4.05
CEC cmolc/kg	398	1.61	88	13.82	14.94
pH	349	3.93	9.16	5.96	1.20
Phophorus mg/kg	406	0.05	547.20	10.73	33.13
OC %	650	0.05	99.17	9.32	8.96
K (s) mm.hr-1	84	104.45	12000	1165.74	2019.81
Valid N (listwise)	7				

6.4 Conclusion

Soil characterisation of the NWU soil database was performed using the populated and quality-controlled soil data collected from different source data. Characterisation was performed according to geographical, morphological, physical, chemical, and hydrological analysis of the soil profiles. From geographical and morphological analysis, it was discovered that soil profiles in the NWU soil database were collected from all nine provinces in South Africa and characterised by a total of 58 soil forms of the South African Classification respectively. From physical, chemical, and hydrological analysis, it was discovered that values for most of the soil properties (sand %, clay %, Bulk density, exchangeable cations, CEC and Phosphorus, OC, pH) encompassed all classes of Hazelton and Murphy (2007). However, values recorded for silt% and K(s), did not encompass all classes of Hazelton and Murphy (2007).

6.5 References

Castrignanó, A., Giugliarini, L., Risaliti, R. & Martinelli, N. 2000. Study of spatial relationships among some soil physico-chemical properties of a field in central Italy using multivariate geostatistics. *Geoderma,* 97: 39-60. https://doi.org/10.1016/S0016-7061(00)00025-2

Hazelton, P. & Murphy, B. 2007. Interpreting soil test results: What do all the numbers mean? CSIRO PUBLISHING.

Heil, K. & Schmidhalter, U. 2012. Characterisation of soil texture variability using the apparent soil electrical conductivity at a highly variable site. *Computer & Geosciences*, 39: 98-110. https://doi.org/10.1016/j.cageo.2011.06.017

Soil Classification Working Groups. 1991. Soil Classification: A Taxonomic System for South Africa. Pretoria: Department of Agricultural Development.

CHAPTER 7 CONCLUSION

There is a need to develop a soil database which captures the available spatially variable soil properties in a format that will enable land users in South Africa to access quality assessed soil data to manage soils sustainably and productively. The aim of this project was to create an NWU soil point database which is standardised, quality-controlled and easily manageable and accessible. This aim was achieved by meeting the following objectives:

The **first objective** was to determine the optimal structure of the NWU soil point database, by examining the structure of an International (WoSIS) and a national (ARC-ISCW) soil databases. This objective was met through the SWOT analysis of the structures of these soil databases, to evaluate the strengths, weaknesses, opportunities, and threats thereof. These factors were evaluated through the observation of their effects on the legibility, data quality, data usage, accessibility, and standardisation. The SWOT analysis results revealed that the legibility of a good soil database is dependent on the ease for data importation and exportation between different databases. Furthermore, legibility is dependent on soil data presentation and comprehensibility. The data quality of a database is dependent on the provision of information about the database, this information includes the definitions of the terminology and units and methods of measurements used in the database. Data usage of the database is dependent on the ease to read, edit, analyse, query, import data, navigate through the database to find information, and perform quality control on the data. Furthermore, data usage is dependent on the ease to understand the software used to store and record data and the language, codes, and classification used in the database. The accessibility is dependent of the ease to obtain the database from the source and the standardisation of a database is dependent on the availability of units and methods of measurements used in the database. The WoSIS database's strengths and opportunities include the ability to combine several TSV files used to record data into a single file and the availability of link tables to facilitate easy navigation through the database. The database is freely available, and its utilization may increase if smaller databases were created for specific geographic areas. The WoSIS database has weaknesses and threats, including the fact that it is an international database used to record vast amounts of soil data collected all over the world. Furthermore, the tables used to record soil properties do not include any units of measurements and first-time users may be deterred due to unfamiliarity with SQL software and TSV files used. The ARC-ISCW database's strengths and opportunities include being hosted in MS Access and can be exported into MS Excel, a more user-friendly software. The weaknesses and threats of the database include the fact that MS Access can only be used by skilled personnel who may resign, resulting in the limited use of the database. Furthermore, the soil properties and methods used for measurements are not recorded in the same table and because the database is unavailable, its potential use is limited. The NWU soil database was developed to include the strengths and opportunities gathered from the analysis of the two databases to improve the legibility, data quality, data usage, accessibility, and standardisation of the database while the weaknesses and threats were negated. Therefore, the NWU soil database was created in an Excel spreadsheet, Excel promotes the ease of data importation and exportation with the “copy” and “paste” option, without distorting the original structure of the database. Furthermore, the ability to record data in table format with soil attribute recorded in designated columns and values in rows results in comprehensible data presentation. Additionally, Excel promotes the ease to edit, analyse, query, and navigate through the database to find information, and perform quality control. The “Attribute Description” worksheet created in the NWU soil database provides information about the data recorded in the database including definition of terminology and symbols used, units and methods of measurements and the data source references. The NWU soil database was used to record South African soil data, therefore, the language, terminology, codes, and classifications used are familiar to the South African Soil Scientists, researchers, and land users. Some of the data recorded in the database will be made available upon request from the NWU. Furthermore, the availability of units and methods of measurements promotes standardisation and performance of quality control on the database.

The **second objective** was to collect soil from a large variety of sources and geographic locations and populate the database structure created under objective 1. This objective was met through the collection of soil data from three academic transcripts, 19 Reports, three Research Reports and two databases. These resulted in a total of 539 soil profiles with 1518 soil horizons recorded in the NWU soil database. Upon data population, data legibility was evaluated, and criterion was used for inclusion of soil data into the database. The data source had to provide “Name of provider” of the data and at least one soil attribute. Furthermore, standardisation was carried out, where the units of measurements from source data were converted to the standard unit used in the NWU soil database.

The **third objective** was to determine and perform quality control measures to ensure the integrity of the NWU soil point database. This objective was met through two stages of data quality control, basic quality control was done by detecting outliers based on the criteria adopted from the WISE 3 database and the African Soil Profile Database version 1.1, and the outlier detection from the creation of box plots in SPSS. Values detected as outliers from outlier detection were flagged and further analysis was done on the values. The analyses involved the revision of detected values to ensure that the values were copied accurately, followed by the analysis of the values based on the surrounding data, comparison of values from values observed in literature. Furthermore, the values were analysed based on the possible impact of terrain or geology. From basic quality control a total of 27 outliers were detected from the database from soil attributes including geographic co-ordinates, sum of particle size, and CEC. These outliers were omitted from the database. Values detected as outliers from outlier detection with boxplots were not omitted from the database because in each case of further analysis it was deemed that the flagged values were likely to occur based on the five-step verification. Therefore, after quality control the NWU soil database was deemed to be a sufficient and accurate source of soil data.

The **fourth objective** was to supply guidelines by which data could be added to the database in future, this objective was met in chapters 4 and 5. For data to be added in the NWU soil database, the criteria set fourth for data population, data legibility evaluation and data standardisation must be taken into consideration.

Furthermore, the data must be subjected to quality control as stipulated in chapter 5. Therefore, the guideline that must be followed is as follows:

Step1 (Supply of database): The database must be requested from the NWU**-** and provided by the relevant supplier.

Step 2 (Data legibility): Data must include information such as “Name of provider” and at least one soil attribute.

Step 3 (Data standardisation): Data must be converted to standard units of measurements used in the NWU soil database.

Step 4 (Data quality control): Basic quality control and outlier detection must be carried out based on criterion set forth for the NWU soil database.

The **final objective** was to characterise the soils from the developed database using measured soil attributes. The objective was met through geographical characterisation of soil data using ArcMap to create a map indicating the location of data points. Morphological characterisation was achieved through the creation of a graph indicating the number of soil forms recorded in the database. Soils in the NWU soil database were collected from all nine provinces in South Africa and contained by 58 of the 73 soil forms of South African classification. Physical, chemical, and hydrological characterisation of soil was achieved with descriptive statistics, using the criteria as adopted from Hazelton and Murphy (2007). Physical properties characterised include soil texture and bulk density, chemical properties include exchangeable cations, Cation Exchange Capacity, Phosphorus status, pH, Organic Carbon, and saturated hydraulic conductivity as a hydrological property. All the physical and chemical properties, excluding silt % are believed to be representative of soils in general as data entries were found in all classes of Hazelton and Murphy (2007), from the lowest to the highest values. Silt % values fit within the very low and high classes and K(s) values which represents the hydrological soil properties represented in the database entries were only found to fit within the high and very high classes of Hazelton and Murphy (2007).

7.1. References

Hazelton, P. & Murphy, B. 2007. Interpreting soil test results: What do all the numbers mean? CSIRO PUBLISHING.

ANNEXURES

Annexure A: WoSIS database structure

wosis_201909_attributes - Notepad

File Edit Format View Help

```
code	type	attribute	unit	profiles	layers	description	accuracy
BDFI33	Horizon	Bulk density fine earth - 33 kPa	kg/dm³	14924	78215	Bulk density of the fine earth fraction*, equilibrated at 33 kPa	35.0
BDFIAD	Horizon	Bulk density fine earth - air dry	kg/dm³	1786	8471	Bulk density of the fine earth fraction*, air dried	35.0
BDFIFM	Horizon	Bulk density fine earth - field moist	kg/dm³	5279	14219	Bulk density of the fine earth fraction*, field moist	35.0
BDFIOD	Horizon	Bulk density fine earth - oven dry	kg/dm³	25124	122693	Bulk density of the fine earth fraction*, oven dry	35.0
BDWS33	Horizon	Bulk density whole soil - 33 kPa	kg/dm³	26268	154901	Bulk density of the whole soil including coarse fragments, equilibrated at 33 kPa	35.0
BDWSAD	Horizon	Bulk density whole soil - air dry	kg/dm³	0	0	Bulk density of the whole soil including coarse fragments, air dried	35.0
BDWSFM	Horizon	Bulk density whole soil - field moist	kg/dm³	0	0	Bulk density of the whole soil including coarse fragments, field moist	35.0
BDWSOD	Horizon	Bulk density whole soil - oven dry	kg/dm³	14588	75422	Bulk density of the whole soil including coarse fragments, oven dry	35.0
CECPH7	Horizon	Cation exchange capacity - buffered at pH7	cmol(c)/kg	54278	295688	Capacity of the fine earth fraction* to hold exchangeable cations, estimated by buffering the
CECPH8	Horizon	Cation exchange capacity - buffered at pH8	cmol(c)/kg	6422	23691	Capacity of the fine earth fraction* to hold exchangeable cations, estimated by buffering the
CFAO	Site	Soil classification FAO unitless	23890	0	Classification of the soil profile according to specified edition (year) of the FAO-Unesco Legend, up to soil unit le
CFGR	Horizon	Coarse fragments gravimetric total	g/100g	39527	203083	Gravimetric content of coarse fragments* in the whole soil	20.0
CFVO	Horizon	Coarse fragments volumetric total	cm³/100cm³	45918	235002	Volumetric content of coarse fragments* in the whole soil	30.0
CLAY	Horizon	Clay total	g/100g	141640	607861	Gravimetric content of < X mm soil material in the fine earth fraction* (e.g. X = 0.002 mm as specified in the analytical method desc
CSTX	Site	Soil classification Soil taxonomy	unitless	21314	0	Classification of the soil profile according to specified edition (year) of USDA Soil Taxonomy, up to
CWRB	Site	Soil classification WRB unitless	26664	0	Classification of the soil profile according to specified edition (year) of the World Reference Base for Soil Resourc
DSDS	Site	Depth of soil - sampled cm	196381	0	Maximum depth of soil described and sampled (calculated)
ECEC	Horizon	Effective cation exchange capacity	cmol(c)/kg	31708	132922	Capacity of the fine earth fraction* to hold exchangeable cations at the pH of the soil (ECEC). Conve
ELCO20	Horizon	Electrical conductivity - ratio 1:2	dS/m	8010	44596	Ability of a 1:2 soil water extract to conduct electrical current	10.0
ELCO25	Horizon	Electrical conductivity - ratio 1:2.5	dS/m	3313	15134	Ability of a 1:2.5 soil water extract to conduct electrical current	10.0
ELCO50	Horizon	Electrical conductivity - ratio 1:5	dS/m	23093	90944	Ability of a 1:5 soil water extract to conduct electrical current	10.0
ELCOSP	Horizon	Electrical conductivity - saturated paste	dS/m	19434	73517	Ability of a water saturated soil paste to conduct electrical current (ECe)	10.0
NITKJD	Horizon	Total nitrogen (N)	g/kg	65356	216362	The sum of total Kjeldahl nitrogen (ammonia, organic and reduced nitrogen) and nitrate-nitrite	10.0
ORGC	Horizon	Organic carbon	g/kg	110856	471301	Gravimetric content of organic carbon in the fine earth fraction*	15.0
PHAQ	Horizon	pH H2O	unitless	130986	613322	A measure of the acidity or alkalinity in soils, defined as the negative logarithm (base 10) of the activity of hydronium ions (H+) i
PHCA	Horizon	pH CaCl2	unitless	66921	314230	A measure of the acidity or alkalinity in soils, defined as the negative logarithm (base 10) of the activity of hydronium ion
PHKC	Horizon	pH KCl	unitless	32920	150447	A measure of the acidity or alkalinity in soils, defined as the negative logarithm (base 10) of the activity of hydronium ions (H+) i
PHNF	Horizon	pH NaF	unitless	4978	25448	A measure of the acidity or alkalinity in soils, defined as the negative logarithm (base 10) of the activity of hydronium ions (H+) i
PHPBYI	Horizon	Phosphorus (P) - Bray I	mg/kg	10735	40486	Measured according to the Bray-I method, a combination of HCl and NH4 F to remove easily acid soluble P forms, largely Al- an
PHPMH3	Horizon	Phosphorus (P) - Mehlich 3	mg/kg	1446	7242	Measured according to the Mehlich-3 extractant, a combination of acids (acetic [HOAc] and nitric [HNO3]), salts (ammo
PHPOLS	Horizon	Phosphorus (P) - Olsen	mg/kg	2162	8434	Measured according to the P-Olsen method:  0.5 M sodium bicarbonate (NaHCO3) solution at a pH of 8.5 to extract P from calcar
PHPRTN	Horizon	Phosphorus (P) - retention	mg/kg	4636	23917	Retention measured according to the New Zealand method	20.0
PHPTOT	Horizon	Phosphorus (P) - total	mg/kg	4022	12976	Determined with a very strong acid (aqua regia and sulfuric acid/nitric acid)	15.0
PHPWSL	Horizon	Phosphorus (P) - water soluble	mg/kg	283	1242	Measured in 1:x soil:water solution (mainly determines P in dissolved forms)	15.0
SAND	Horizon	Sand total	g/100g	105547	491810	Larger than Y mm fraction of the fine earth fraction*; Y as specified in the analytical method description (e.g. Y = 0.05 mm to 2 mm*
SILT	Horizon	Silt total	g/100g	133938	575913	X to Y mm fraction of the fine earth fraction*; Y as specified in the analytical method description (e.g. Y = 0.05 mm)	15.0
TCEQ	Horizon	Calcium carbonate equivalent total	g/kg	51991	222242	The content of carbonate in a liming material or calcareous soil calculated as if all of the carbonate is in
TOTC	Horizon	Total carbon (C)	g/kg	32662	109953	Gravimetric content of organic carbon and inorganic carbon in the fine earth fraction*	10.0
WG0006	Horizon	Water retention gravimetric - 6 kPa	g/100g	863	4264	Soil moisture content by weight, at tension 6 kPa (pF 1.8)	20.0
WG0010	Horizon	Water retention gravimetric - 10 kPa	g/100g	3357	14739	Soil moisture content by weight, at tension 10 kPa (pF 2.0)	20.0
WG0033	Horizon	Water retention gravimetric - 33 kPa	g/100g	21116	96354	Soil moisture content by weight, at tension 33 kPa (pF 2.5)	20.0
```

Figure A1- 1: Figure displaying the structural format of the wosis_201909_attributes tsv. file.

wosis_201909_profiles - Notepad

File Edit Format View Help

```
profile_id	dataset_id	country_id	country_name	geom_accuracy	latitude	longitude	dsds	cfao_version	cfao_major_group_code	cfao_major_group	cfao_
36897	{BE-UplandsI}	BE	Belgium	1e-06	50.64988900	4.66690100	100
36898	{BE-UplandsI}	BE	Belgium	1e-06	50.58396200	4.46211400	97
36899	{BE-UplandsI}	BE	Belgium	1e-06	50.59787600	4.68760700	109
36900	{BE-UplandsI}	BE	Belgium	1e-06	50.63359900	4.68178300	94
36901	{BE-UplandsI}	BE	Belgium	1e-06	50.62320400	4.46603500	100
36902	{BE-UplandsI}	BE	Belgium	1e-06	50.61051700	4.61912800	103
36903	{BE-UplandsI}	BE	Belgium	1e-06	50.59850500	4.77279800	103
36904	{BE-UplandsI}	BE	Belgium	1e-06	50.59607900	4.86300900	94
36905	{BE-UplandsI}	BE	Belgium	1e-06	50.54104100	4.64208500	106
36906	{BE-UplandsI}	BE	Belgium	1e-06	50.54647300	4.60626400	100
36907	{BE-UplandsI}	BE	Belgium	1e-05	50.53257300	4.76941000	109
36908	{BE-UplandsI}	BE	Belgium	1e-06	50.81454200	4.59723900	310
36909	{BE-UplandsI}	BE	Belgium	1e-06	50.81431500	4.59691300	500
36910	{BE-UplandsI}	BE	Belgium	1e-05	50.81357000	4.59547600	355
36911	{BE-UplandsI}	BE	Belgium	1e-06	50.76083900	4.73591100	500
36912	{BE-UplandsI}	BE	Belgium	1e-06	50.60442700	4.72296400	106
36913	{BE-UplandsI}	BE	Belgium	1e-06	50.67673700	4.70761400	82
36914	{BE-UplandsI}	BE	Belgium	1e-06	50.71306400	4.90136400	109
36915	{BE-UplandsI}	BE	Belgium	1e-06	50.65953800	4.74132500	109
36916	{BE-UplandsI}	BE	Belgium	1e-06	50.60357300	4.73553100	109
36917	{BE-UplandsI}	BE	Belgium	1e-06	50.65754700	4.74175800	109
36918	{BE-UplandsI}	BE	Belgium	1e-05	50.70681000	4.73383800	97
36919	{BE-UplandsI}	BE	Belgium	1e-06	50.75764700	4.97271800	109
36920	{BE-UplandsI}	BE	Belgium	1e-06	50.61330300	4.96577200	109
36921	{BE-UplandsI}	BE	Belgium	1e-06	50.59045400	4.45890500	109
36922	{BE-UplandsI}	BE	Belgium	1e-06	50.74797700	4.95401400	97
36923	{BE-UplandsI}	BE	Belgium	1e-06	50.75325300	4.73923900	100
36924	{BE-UplandsI}	BE	Belgium	1e-05	50.53751000	4.76084100	109
36925	{BE-UplandsI}	BE	Belgium	1e-06	50.56286800	4.65653300	109
36926	{BE-UplandsI}	BE	Belgium	1e-06	50.65181800	4.98069600	106
36927	{BE-UplandsI}	BE	Belgium	1e-06	50.77590800	4.70283800	109
36928	{BE-UplandsI}	BE	Belgium	1e-06	50.63503200	4.74002400	106
36929	{BE-UplandsI}	BE	Belgium	1e-05	50.60274000	4.94020200	109
36930	{BE-UplandsI}	BE	Belgium	1e-06	50.64657900	4.89337200	100
36931	{BE-UplandsI}	BE	Belgium	1e-06	50.61909500	4.49838800	109
36932	{BE-UplandsI}	BE	Belgium	1e-06	50.67004600	4.91273700	85
36933	{BE-UplandsI}	BE	Belgium	1e-06	50.78416600	4.94271200	109
36934	{BE-UplandsI}	BE	Belgium	1e-06	50.65777600	4.94415700	103
36935	{BE-UplandsI}	BE	Belgium	1e-06	50.68462800	4.88376600	85
36936	{BE-UplandsI}	BE	Belgium	1e-06	50.59419700	4.68434200	106
36937	{BE-UplandsI}	BE	Belgium	1e-06	50.63617800	4.94545100	109
```

Ln 1, Col 1 100% Unix (LF) UTF-8

Figure A1- 2: Figure displaying the structural format of the wosis_201909_profiles tsv. file.

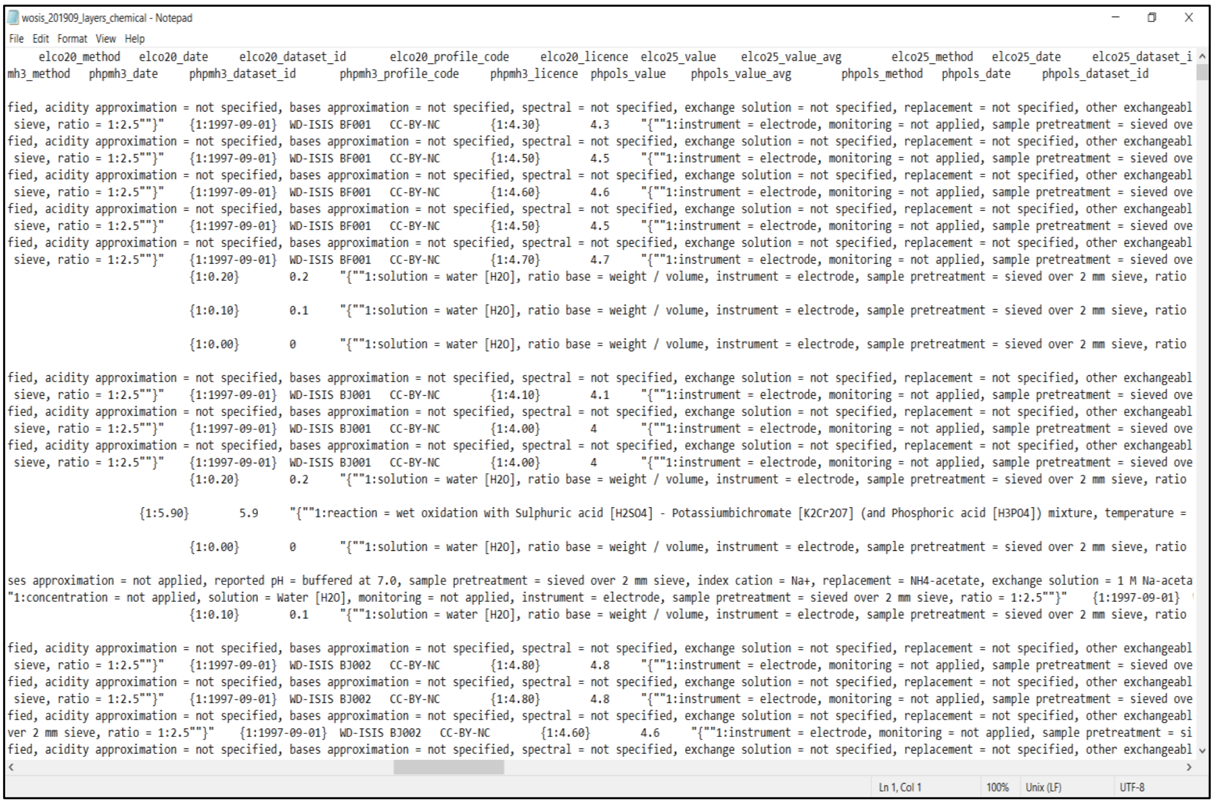


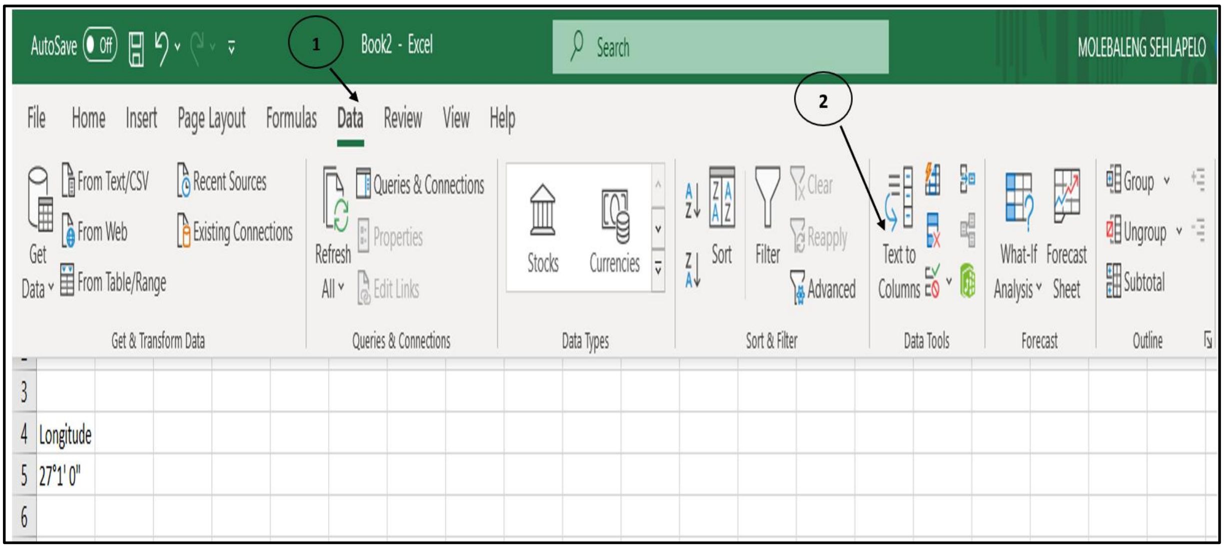
Figure A1- 3: Figure displaying the structural format of the wosis_201909_chemical tsv. file.

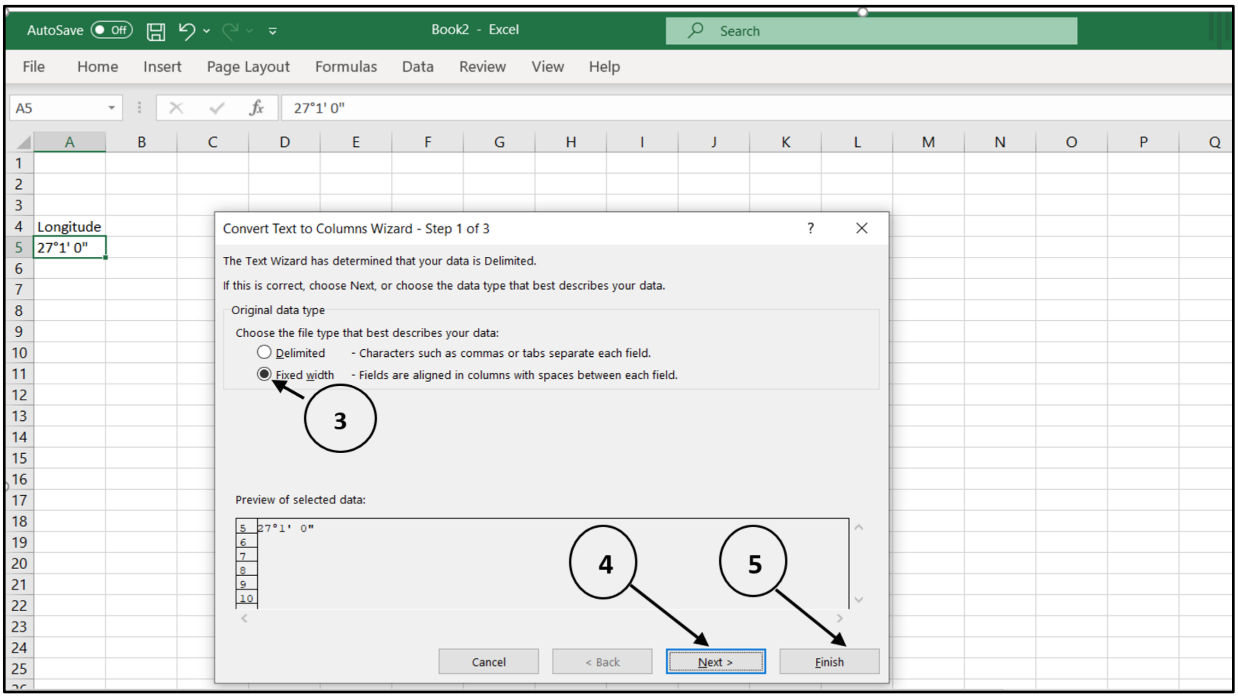
Figure A1- 4: Figure displaying the structural format of the wosis_201909_physical tsv. file.

Annexure B: Converting DSM to DD in Excel

Split the DMS:

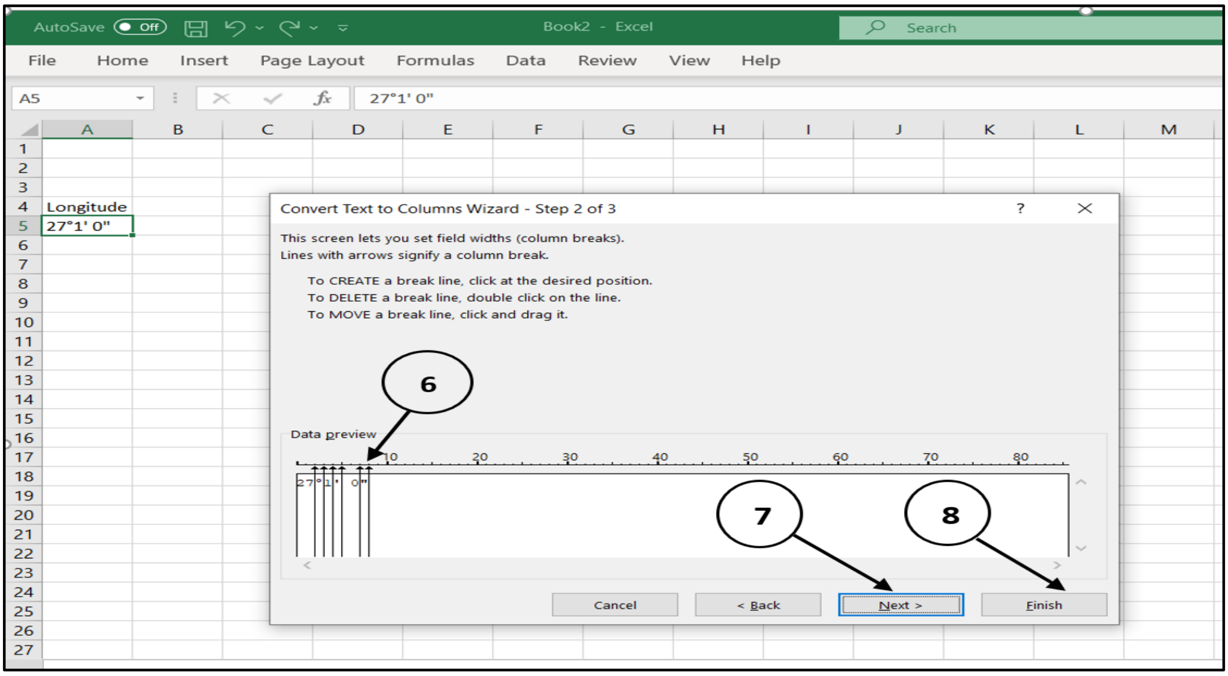
- Select the data - coordinates to be converted
- Select ‘data’-**step 1** on tab then ‘Text to columns’ from ribbon-**step 2**
- From the wizard that will appear preceding the previous steps, choose the ‘fixed width’ option from the ‘next’ then finish-**steps 3,4,5** this option will ensure that the selected data is separated between fields and the alignment of the field in columns.





- The preceding wizard will give an option of creating break lines, which will represent the different columns each feature will apper in. The figures representing the degrees, minutes and seconds, should be seperated from the symbols of measurements-**step 6**

- To complete this step, the ‘next’ and ‘finish’options should be selected- **steps 7,8**.



Calculations:

- After the completion of the steps from the ‘convert text to column wizard’, the longitude value will be separated into columns representing the values and symbols as selected from the creation of the break lines-**step 9**

-From this the columns representing the symbols should be deleted as they will not be used in calculation (Table 4.1b), the column left will show the degrees, minutes and seconds values respectively.

- The calculation of DMS to DD is done by using the equations in excel, where the degrees value is added to the minutes value divided by 60 minutes and added to the minutes value divided by 3600 second- D+M/60+S/3600- **step 10** . Table 4.1c represents the final answer in DD-**step 11**.

Annexure C: Outlier values detected using boxplots

Table C1-1: Outliers detected from basic quality control.

Source	Row no.	Clay %	coSand %	meSand %	fiSand %	vfiSand %	coSilt %	fSilt %	Sum
HOSASH	73	24,00	8,00	10,00	18,00	16,00	115,00	9,00	200,00
HOSASH	103	18,77	29,75	18,05	15,30	7,65	10,13	11,35	111,00
HOSASH	111	18,00	29,00	15,00	16,00	8,00	8,00	18,00	112,00
HOSASH	112	27,00	13,00	15,00	16,00	9,00	12,00	27,00	119,00
HOSASH	113	30,00	14,00	15,00	16,00	10,00	8,00	30,00	123,00
HOSASH	144	30,60	26,80	9,90	27,20	13,30	8,70	14,30	130,80
HOSASH	145	32,67	11,90	18,70	18,80	10,40	10,60	19,70	122,78
HOSASH	206	22,60	6,40	29,60	52,00	0,70	31,20	21,80	164,30
HOSASH	218	39,50	11,30	11,10	27,90	36,40	76,90	21,70	224,80
HOSASH	222	57,80	4,30	5,70	4,30	5,10	87,10	44,90	209,20
HOSASH	223	2,20	53,00	1,90	37,70	29,40	33,70	1,20	159,10
HOSASH	227	44,90	2,80	3,30	3,00	48,10	53,20	81,40	236,70
HOSASH	231	36,10	1,98	27,80	35,23	46,65	25,65	37,65	211,06
HOSASH	238	38,00	25,30	3,50	10,60	21,70	22,00	40,00	161,10

Table C1-2: Outliers detected from basic quality control.

CEC 1-120cmol/kg		
Source	**Row no.**	**Outliers**
HOSASH	114	252,1739
HOSASH	117	573,913
HOSASH	118	547,8261
HOSASH	125	156,5217
HOSASH	126	1017,391
HOSASH	171	154,7826
HOSASH	172	152,1739

Table C1-3: Outliers values detected from boxplots.

Source	Case no.	Mild Outlier Values	Source	Case no.	Extreme Outlier Values
Clay %					
Van Tol,	1078	68,7			
coSand %					
HOSASH	101	37,2	HOSASH	223	53
HOSASH	105	46			
HOSASH	107	38			
HOSASH	118	32,3			
HOSASH	125	36,4			
HOSASH	126	33			
HOSASH	216	36,9			
HOSASH	228	37			
meSand %					
HOSASH	182	43,7	HOSASH	194	74,7
HOSASH	196	52,1	HOSASH	199	69
HOSASH	210	42,9	HOSASH	211	76,6
HOSASH	213	56,3			
fiSand %					
ARC-ISCW	13	63,4			
ARC-ISCW	18	58,8			
vfiSand %					
HOSASH	218	36,4			
HOSASH	227	48,1			
HOSASH	231	46,65			
Manyevere A.	242	41,19			
Manyevere A.	264	41,19			
Manyevere A.	267	42,69728			
Manyevere A.	270	41,19			
Manyevere A.	276	37,96602			
Manyevere A.	284	48,79702			
Manyevere A.	307	41,19			
coSilt %					
HOSASH	92	21	HOSASH	73	115
HOSASH	231	25,65	HOSASH	206	31,2
			HOSASH	218	76,9
			HOSASH	222	87,1
			HOSASH	223	33,7
			HOSASH	227	53,2
fiSilt %					
HOSASH	112	27	HOSASH	222	44,9
HOSASH	149	27	HOSASH	227	81,4
HOSASH	172	27	HOSASH	238	40
HOSASH	177	27			
HOSASH	231	37,65			
HOSASH	237	26,5			

Table C1-4: Outliers values detected from boxplots.

NH4OAc(Na) (cmol/kg)					
Source	**Case No.**	**Mild outlier Values**	**Source**	**Case No.**	**Extreme outlier Values**
HOSASH	118	1,1826087	ARC-ISCW	10	4,5
HOSASH	128	1,77	HOSASH	101	1,913043478
DSA	554	1,36	HOSASH	108	2,7
DSA	555	1,85	HOSASH	114	8,347826087
DSA	559	1,45	HOSASH	137	6,1913
DSA	560	1,89	HOSASH	138	8,27826
DSA	573	1,35	HOSASH	140	6,81739
DSA	598	1,8	HOSASH	141	6,81739
			HOSASH	144	7,3
			HOSASH	148	6,95652
			HOSASH	150	10,9913
			HOSASH	152	16,62609
			HOSASH	164	6,6087
			HOSASH	165	4,5913
			HOSASH	170	10,85217
			HOSASH	171	12,31304
			HOSASH	172	10,78261
			HOSASH	174	6,33043
			HOSASH	175	6,26087
			HOSASH	178	8,48696
			DSA	454	4,308695652
			DSA	552	2,78
			DSA	564	12,55
			DSA	565	15,01
			DSA	574	2,74
			DSA	583	8,55
			DSA	599	2,8
NH4OAc (K) (cmol/kg)					
ARC-ISCW	57	1,3	HOSASH	105	1,435897436
ARC-ISCW	100	1,1794872	HOSASH	108	3,08
ARC-ISCW	152	1,23	HOSASH	109	2,1
HOSASH	171	0,94	HOSASH	110	3,9
NLEIP	357	0,93	HOSASH	137	2,13
DSA	496	0,9461538	HOSASH	147	2,87
DSA	501	0,9205128	HOSASH	165	1,93
			HOSASH	168	5,5
			HOSASH	170	3,65
			HOSASH	172	1,48
			HOSASH	174	1,97
			HOSASH	175	4,14
			HOSASH	176	2,38
			HOSASH	178	3,08
			Manyevere A.	310	1,341025641
			DSA	468	2,030769231
			DSA	513	1,664102564

Table C1-5: Outliers values detected from boxplots.

Source	Case No.	Mild outlier Values	Source	Case No.	Extreme outlier Values
Mehlich 3 (Na) (mg/kg)					
			DSA	699	562
			DSA	700	773
			DSA	711	658
			DSA	725	630
			DSA	729	427
			DSA	728	400
			DSA	734	716
Mehlich 3 (K) (mg/kg)					
DSA	684	223	DSA	721	303
Mehlich 3 (Ca) (mg/kg)					
DSA	737	3463	DSA	488	6496
Mehlich 3 (Mg) (mg/kg)					
DSA	694	1209	DSA	695	1912
DSA	699	1528		728	1895
DSA	700	1364		734	2343
DSA	725	1663			
DSA	729	1841			
di-ammonium EDTA (Zn) (mg/kg)					
ARC-ISCW	21	0,92	ARC-ISCW	16	1,38
ARC-ISCW	56	0,88	ARC-ISCW	30	2,21
ARC-ISCW	62	1,08	ARC-ISCW	36	1,53
			ARC-ISCW	44	3,23
			ARC-ISCW	50	1,77
di-ammonium EDTA (Mn) (mg/kg)					
ARC-ISCW	37	127	ARC-ISCW	9	578,3
ARC-ISCW	50	106,3	ARC-ISCW	10	784,7
			ARC-ISCW	16	196,6
			ARC-ISCW	17	148,8
			ARC-ISCW	18	342,2
			ARC-ISCW	20	216,6
di-ammonium EDTA (Cu) (mg/kg)					
ARC-ISCW	9	3,86			
ARC-ISCW	50	9,91			
ARC-ISCW	52	7,11			
di-ammonium EDTA (Co) (mg/kg)					
ARC-ISCW	50	20,22			
ARC-ISCW	51	19,2			
ARC-ISCW	52	20,67			
di-ammonium EDTA (B) (mg/kg)					
			ARC-ISCW	37	2,45

Table C1-5: Outliers values detected from boxplots.

DTPA (Zn) (mg/kg)					
Source	Case No.	Mild Outlier Values	Source	Case No.	Extreme outlier values
UFH	355	2,29	UFH	331	8,2
UFH	423	3,61	UFH	356	36,11
			UFH	357	42,95
			UFH	358	27,22
			UFH	370	17,71
			UFH	402	21,26
			UFH	403	5,44
			UFH	405	50,58
			DSA	597	6,4
			DSA	598	8,4
			DSA	599	10,4
			DSA	600	12,4
			DSA	601	14,4
			DSA	602	2,4
			DSA	603	4,4
			DSA	604	16,4
			DSA	605	18,4
DTPA (Mn) (mg/kg)					
UFH	343	104,68	UFH	338	174,02
UFH	351	145,23	UFH	345	256,03
UFH	356	118,28	UFH	358	169,89
UFH	357	151,87	UFH	370	368,82
UFH	365	107,07	UFH	371	187,58
UFH	374	111,18	UFH	375	265,52
UFH	378	103	UFH	413	307,74
UFH	379	100,27			
UFH	397	148,99			
DTPA (Cu) (mg/kg)					
UFH	374	8,53	UFH	356	10,53
UFH	375	7,37	UFH	405	29,2
DSA	597	7	DSA	599	11
DSA	598	9	DSA	600	13
			DSA	601	15
			DSA	605	17
			DSA	604	19
DTPA (B) (mg/kg)					
UFH	353	0,21	UFH	405	0,17
UFH	357	0,15	UFH	418	0,18
UFH	359	0,18	UFH	423	1,39
UFH	399	0,03	DSA	597	6,2
			DSA	598	8,2
			DSA	599	10,2
			DSA	600	12,2
			DSA	601	14,2
			DSA	602	2,2
			DSA	603	4,2
			DSA	604	16,2
			DSA	605	18,2

Table C1-7: Outliers values detected from boxplots.

Mehlich 3 (Zn) (mg/kg)					
Source	**Case No.**	**Mild Outlier Values**	**Source**	**Case No.**	**Extreme Outlier Values**
DSA	684	54,88	DSA	693	1777,08
DSA	728	6,76	DSA	694	432,56
DSA	733	46,6	DSA	730	672,25
			DSA	731	123,29
			DSA	736	321,19
			DSA	737	697,93
			DSA	739	763,41
Mehlich 3 (Mn) (mg/kg)					
DSA	706	232,22	DSA	694	471,45
DSA	727	264			
Mehlich 3 (Cu) (mg/kg)					
DSA	727	7,19	DSA	730	12,02
DSA	733	7,53			
Mehlich 3 (B) (mg/kg)					
DSA	678	1,58	DSA	684	2,97
DSA	693	1,96			
ISFEI P (mg/kg)					
ARC-ISCW	16	2,7	ARC-ISCW	9	6,8
			ARC-ISCW	24	36,31
Mehlich 3 P (mg/kg)					
DSA	482	18			
DSA	684	18			
DSA	705	17			
DSA	730	22			
Bray 1 P (mg/kg)					
UFH	404	23,8	UFH	356	547,2
DSA	431	32	UFH	370	133,8
DSA	598	23,6	UFH	402	193,6
DSA	601	24	UFH	403	67,6
DSA	605	24	UFH	404	282,8
			DSA	555	46,38
			DSA	602	41,8
Bray 2 P (mg/kg)					
DSA	543	30			
DSA	547	34			
DSA	693	32			
DSA	730	30			
CBD (Fe) %					
ARC-ISCW	51	6,11			
ARC-ISCW	52	5,65			
CBD (Al) %					
	17	0,31	ARC-ISCW	50	0,6
	18	0,29			
		0,29			
CBD (Mn) %					
ARC-ISCW	47	0,24	ARC-ISCW	54	0,18
ARC-ISCW	48	0,37	ARC-ISCW	58	0,48

Table C1-8: Outliers values detected from boxplots.

CEC (cmol(+)/kg)					
Source	**Case No.**	**Mild Outlier Values**	**Source**	**Case No.**	**Extreme Outlier Values**
HOSASH	108	80	HOSASH	114	252,173913
HOSASH	110	43,478261	HOSASH	117	573,9130435
HOSASH	150	42,6087	HOSASH	118	547,826087
HOSASH	232	64	HOSASH	125	156,5217391
HOSASH	612	12	HOSASH	126	1017,391304
			HOSASH	147	70,43478
			HOSASH	152	142,6087
			HOSASH	171	154,78261
			HOSASH	172	152,17391
			HOSASH	231	88
Saturated Paste (EC) (mS/m)					
DSA	498	88,3	DSA	499	106,1
DSA	432	74,1			
DSA	454	69,3			
EC (mS/m)					
DSA	517	119,7	DSA	554	198,9
DSA	518	116,8	DSA	560	202,7
DSA	555	114,1	DSA	563	187,9
			DSA	564	1712
			DSA	573	280
			DSA	574	1123
			DSA	579	178,5
				582	632
				883	1041
				599	260,8
				605	128,5
WBM-Black (OC) %					
HOSASH	168	3,9	ARC-ISCW	10	0,2
UFH	325	4,02	UFH	351	12,47
UFH	356	4,02	UFH	355	14,78
UFH	368	3,98			
UFH	376	4,1			
UFH	385	4,27			
UFH	396	4,27			
OC %					
Van Tol	1049	2,4	Van Tol	1040	3,46
Van Tol	1051	2,1	Van Tol	1043	3,5
Van Tol	1052	0,4	Van Tol	1088	9,36
Van Tol	1096	2,32	Van Tol	1089	5,47
			Van Tol	1092	4,45
			Van Tol	1095	3,59

Table C1-9: Outliers values detected from boxplots.

Source	Case No.	Mild Outlier Values	Source	Case No.	Extreme Outlier Values
Water retention (-33kPa) %					
ARC-ISCW	9	38,9			
ARC-ISCW	10	41,2			
ARC-ISCW	18	31,3			
ARC-ISCW	36	41,7			
ARC-ISCW	37	44,6			
Water retention (-80kPa) %					
ARC-ISCW	10	35,3			
ARC-ISCW	36	35,5			
ARC-ISCW	37	36,6			
Water retention (-500kPa) %					
ARC-ISCW	37	29,6			
AWR %					
ARC-ISCW	18	87,7	ARC-ISCW	17	273,2
			ARC-ISCW	39	201,7
			ARC-ISCW	55	175,4
K (0) (mm/ hr-1)					
HOSASH	66	335,74	HOSASH	79	494,8
HOSASH	67	372,43			
HOSASH	82	329,9			
HOSASH	115	304,86052			
HOSASH	123	263,03422			
K (30) (mm/ hr-1)					
HOSASH	66	40,84			
HOSASH	67	41,6			
HOSASH	79	5			
K (60) (mm/ hr-1)					
UND	1398	34,43	UND	1210	61,19
UND	1473	32,64	UND	1331	255,93
UND	1504	36	UND	1332	128,41
UND	1506	39,6	UND	1393	60,36
			UND	1418	59,14
			UND	1419	55,25
			UND	1476	40,39
			UND	1477	40,67
			UND	1520	601,25
K (80) (mm/ hr-1)					
HOSASH	82	20,2			
HOSASH	66	21,29			
K (150) (mm/ hr-1)					
HOSASH	165	10,3	UND	1331	69,61
HOSASH	1210	11,19	UND	1332	17,67
UND	1436	11,36	UND	1393	18,27
			UND	1418	23,67
			UND	1419	16,71
			UND	1520	132,9

Table C1-10: Outliers values detected from boxplots.

K(s) (mm/ hr-1)					
Source	**Case No.**	**Mild Outlier Values**	**Source**	**Case No.**	**Extreme Outlier Values**
DSA	740	842,12	Van Tol	1173	8000
DSA	746	842,12	Van Tol	1176	1655,172414
DSA	760	842,12	Van Tol	1178	6857,142857
Van Tol	1165	854,21266	Van Tol	1179	8000
Van Tol	1505	1373,4	Van Tol	1182	4000
			Van Tol	1185	4000
			Van Tol	1186	1655,172414
			Van Tol	1187	6000
			Van Tol	1191	3692,307692
			Van Tol	1194	4000
			Van Tol	1196	12000
			Van Tol	1509	1770,6