A Predictive Model for Determining Vegetable-Derived Macronutrients for a Diabetic Patient

By
Allan Odindo Vikiru
098587

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Information Technology at Strathmore University

School of Computing and Engineering Sciences
Strathmore University
Nairobi, Kenya

15 April 2024

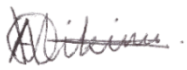
Declaration and Approval

Declaration

I, Allan Odindo Vikiru, declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University

Name: Allan Odindo Vikiru

Signature:  Date: 15 April 2024

Approval

The thesis of Allan Odindo Vikiru was reviewed and approved (for examination) by the following:

7th April, 2024

Dr. Vincent Omwenga
Senior Lecturer, School of Computing and Engineering Sciences
Strathmore University

Dr. Julius Butime
Dean, School of Computing and Engineering Sciences
Strathmore University

Dr. Bernard Shibwabo
Director of Graduate Studies
Strathmore University

Abstract

Current tools for predicting suitable vegetable plants for consumption by patients of type-II diabetes do not factor for their geolocation. This limits them from accessing and consuming vegetables with the required nutrient content that is needed to control the disease. Given soils have a direct influence on the macronutrient levels found in a plant, there was need to predict optimally the amount of nutrients ingested by the patient, given the geolocation, as soils act as a source of nutrients for vegetable plants. Therefore, this study developed a predictive model that enabled referencing of location-specific nutrients through a georeferenced map of soil macronutrients. The model applied was a hybrid geospatial model that is based on multiple linear regression kriging, that combined a geostatistical and a machine learning technique for predictive mapping, to enhance accuracy of prediction of generated maps. This model was compared against others by support vector machines, multiple linear regression, and regression kriging. However, it performed poorer in map generation compared to the other three models, with the model based on support vector machines providing the highest level of accuracy. The maps generated by this model were then applied in a tool for nutritionists to determine suitable vegetables that is supported by soil at a specific location. This tool will greatly assist in providing tailored dietary plans to type-II diabetic patients.

Keywords: soil macronutrients, human nutrition, hybrid geospatial model, support vector machines, plant determinant tool

VT OMNES VNVM SINT

Table of Contents

VT OMNES VNVM SINT

VT OMNES VNVM SINT

List of Figures

VT OMNES VNVM SINT

List of Tables

VT OMNES VNVM SINT

List of Equations

Abbreviations/Acronyms

AfSP – Africa Soil Profiles Database

DSM – digital soil mapping

ExK – exchangeable potassium

ExMg – exchangeable magnesium

GIS – geographical information system

IDW – inverse distance weighting

ML – machine learning

MLR – multiple linear regression

MLRK – multiple linear regression kriging

NACOSTI – National Commission for Science, Technology, and Innovation

NCDs – non-communicable diseases

OK – ordinary kriging

RMSE – root mean squared error

RK – regression kriging

SVM – support vector machine

SU-ISERC – Strathmore University – Institutional Scientific Ethics Review Committee

UML – Unified Modified Language

Definition of Terms

Digital soil mapping – development of spatial soil information systems by use of soil types and properties from a set of soil observations mapped against related environmental variables using a statistical method (Lagacherie, 2008).

Environmental covariates – a set of independent variables that describe the environmental conditions from which a soil and its properties are found and are used to create a relationship between a soil sample and its environment (Heung et al., 2021).

Geostatistics – a branch of applied statistics that determines the degree of autocorrelation between two measured independent points in a space and uses spatial information of a measured point to predict the value of unmeasured points (Kumar & Sinha, 2018).

Hybrid technique – combining a machine learning technique and geostatistical technique into a single approach for spatial modelling. The machine learning technique caters for the deterministic trend while geostatistical approach handles the stochastic trend (Matinfar et al., 2021).

Machine learning – use of a model with a data analysis technique and component of artificial intelligence to identify patterns within a set of data (Khaledian & Miller, 2020).

VT OMNES VNVM SINT

Acknowledgments

I acknowledge God's favour in my life and academic journey so far. I deeply thank my supervisor for this study, Dr. Vincent Omwenga for his support and guidance in developing this thesis. I also extend my appreciation to members of faculty at the School of Computing and Engineering Sciences for their constructive feedback. Lastly, I truly appreciate my family and friends for their never-ending support in my masters' studies.

Chapter 1: Introduction

1.1. Background

Soil fertility is fundamental to human well-being. This is presented through the notion of soil security, that demonstrates the contribution of soil to sustainable development. Soil security is illustrated through five dimensions which cover economic, social, biophysical, and legal aspects on protection, health, and quality of soil (A. McBratney et al., 2014). Two dimensions: capability and condition, directly tackle concerns regarding human nutrition, by looking into the biophysical attributes of soil that contribute towards human health. They focus on the soil's ability and state to produce abundant and nourishing food crops and transmit essential nutrients to animals and humans that feed directly or indirectly from soil-grown foods (Brevik et al., 2017; FAO, 2022).

Brevik et al. (2020) highlights the growing need to understand the roles of soil and plant nutrients in human nutrition. Besides understanding means for achieving optimum soil conditions for crop growth, exploring this notion further demonstrates the contribution of soil nutrients towards improving human health. Macronutrients such as calcium, potassium and sodium in soil, present in crops are essential in curbing the effect of malnutrition and deficiency diseases (Huang et al., 2020). More importantly, diets deficient in these nutrients can lead to contraction of non-communicable diseases (NCDs) such as lack of magnesium resulting in contraction of type-II diabetes (Dubey et al., 2020), which is a growing concern of public health in Kenya (Sarah et al., 2021).

Agronomic biofortification is recognised as a cheap and efficient way of enhancing mineral intake in human diet (Dasgupta et al., 2023; Szerement et al., 2022). This involves agricultural management processes such as applying fertilisers in soil to enhance subsequent mineral intake in crops. Before the correction of micronutrients, soil is first assessed to determine the available properties. In large scale, these assessments are represented as spatial data, which appropriately supports decision making processes in agricultural and land management as well as policy making for sustainability development (Buenemann et al., 2023; Kimsey et al., 2020). Use of soil maps is preferred over other techniques such as spectral analysis (spectroscopy) and remote sensing since they are relatively less time-consuming and require less resources for analysis (Asrat et al., 2023; Pierangeli et al., 2022). Moreover, much as they are inexpensive, soil sensing methods are not as effective for measuring variations of soil properties across geographic regions (Dasgupta et al., 2023).

Generation of soil maps is done by digital soil mapping (DSM), which is an improvement of conventional mapping methods. These techniques involve conceptualising soil models, using images and related maps to determine soil components, plotting the map, and validating the prepared map through field observations (Minasny & McBratney, 2016). Not only is this time-consuming and costly to administer, but it also generates static maps which do not meet the growing demand for output that illustrate relationships between soil attributes and other related factors (Buenemann et al., 2023). DSM improves this by incorporating statistical models that analyse takes in soil data, from georeferenced maps or collected observations, relates it to identified environmental covariates, and produces a spatial soil information system in the form of raster data of predicted values (Heung et al., 2021; Minasny & McBratney, 2016).

1.2. Problem Statement

Management of type-II diabetes requires highly individualised diet planning that takes into account varying factors for a patient such as their location, sociocultural issues, economic conditions and their eating habits (Minari et al., 2023). Availability of food within a particular region can affect the rate of diabetes contraction and management (Hill-Briggs et al., 2020) and consumption in poor quality vegetables characterised by deficiencies in soil nutrients can increase the effect of diabetes in patients (Dias, 2019). Keesstra et al. (2016) and Stewart et al. (2020) highlight a lack of knowledge transfer on soil fertility management to enhance human nutrition, by which Berkhout et al. (2019) suggests the application of soil maps to establish a statistical connection between soil nutrients and human health.

Predictive modelling for DSM largely follows spatial interpolation methods that are based in either machine learning (ML) or geostatistics, that are part of geographic information systems (GIS). In spatial analysis for soil properties, popular ML approaches such as Support Vector Machines (SVM) and regression-based analysis, and geostatistical-based methods such as Ordinary Kriging (OK), are applied to estimate the value of unsampled locations by calculating distances between sampled locations (Folorunso et al., 2023; Ouabo et al., 2020). Each technique, however, when used individually presents some drawbacks such as underfitting (Dasgupta et al., 2023) and non-linearity between variables (Gao et al., 2021), that limit them from achieving reliable accuracies of prediction. Therefore, the recommended technique is a hybrid model that takes advantage of features in both ML and geostatistical-based approaches to overcome pitfalls in individual methods (Folorunso et al., 2023) and fill a present gap in predicting soil macronutrients at a regional scale (Dasgupta et al., 2023).

1.3. Objective of the Study

1.3.1. General Objective

To develop a predictive model that determines vegetable derived macronutrients using georeferenced maps of soil macronutrients to meet the micronutrient requirements in diabetic patients.

1.3.2. Specific Objectives

The specific objectives that guided the research are:

i. To investigate the relationship between soil macronutrients and macronutrients in plant vegetables,
ii. To examine the relationship between plant vegetable macronutrients and the influence on human nutrition,
iii. To analyse the data modelling techniques for generating georeferenced soil maps,
iv. To develop a hybrid geospatial model that generates digital soil macronutrient maps of identified regions in Kenya, and
v. To validate the performance of the developed model in determining the suitable plant based on required macronutrients.

1.4. Research Questions

The research questions that were addressed by the study are:

i. What is the relationship between soil macronutrients and plant macronutrients in plant vegetables?
ii. How do plant vegetable macronutrients influence human nutrition?
iii. What modelling techniques are used in generating georeferenced soil maps?
iv. How was a hybrid geospatial model developed to generate soil macronutrient maps of identified regions in Kenya?
v. How was the developed map tested for uniformity in determining suitable plants based on required macronutrients?

1.5. Justification

Nutrient-deficient soils produce nutrient-deficient plants, and their consumption by people brings about nutritional deficiencies in diet, consequently resulting in poor health. This can be characterised by

contraction of non-communicable diseases that result from poor diet patterns such as type-II diabetes. Management of these diseases involves inclusion of specific nutrients that promote the patient's state of health (Ruthsatz & Candeias, 2020).

As spelt out in the Nutritionists and Dieticians Act (2008), specialists in human diet and nutrition need to be enlightened on suitable nutritional characteristics to offer effective recommendations on diet. To support this notion, this study focused on the contribution of soil nutrients to human nutrition, by comparing them against plants for human consumption that grow from soil nutrients. Understanding soil fertility would ensure proper utilisation and management of soil, which enables effective flow of nutrients from plants that support human health (Stewart et al., 2020).

1.6. Scope and Limitation

This research focused on spatial distribution of two macronutrients: magnesium and potassium on Kenyan soil. This is due to their presence in vegetables consumed across the country, which allowed for uniformity in testing the developed model, and their influence in management of type-II diabetes. The resulting soil maps were prepared using predictive models; thus, it will be crucial for further verification by specialists be done, to ensure their applicability in real-world settings.

VT OMNES VNVM SINT

Chapter 2: Literature Review

2.1. Introduction

This chapter looks into published works that support the study, based on the first, second, and third research objectives in section 1.3.2. It discusses the relationship between soil nutrients and human nutrition, and analyses techniques used for digital soil mapping, which include geostatistical, machine learning and hybrid geospatial techniques used in spatial interpolation of soil properties. Lastly, it details a summary of works closely related to the study, and a conceptual framework for development of the hybrid geospatial model and plant determinant tool.

2.2. Soil Nutrient Analysis for Human Nutrition

2.2.1. Soil and Plant Macronutrients

Soil directly supports the growth of plants, by facilitating the development of their root systems and serving as a source for nutrients and plant-available water (Comerford, 2005). Fageria et al. (2011) defines the ability of plants to take up nutrients as nutrient bioavailability, and the rate of nutrient uptake by different plants is determined by both properties in soil such as soil type and water content, and plants such as root density (Raynaud & Leadley, 2004).

There are six identified macronutrients in soil, each with identified functions in plants. Nitrogen is needed in large quantities to carry out primary functions such as photosynthesis, increasing leaf size and development of fruits and seeds (Johnson & Mirza, 2020; Zewide & Melash, 2021). Phosphorous is essential for plant maturity through processes such as germination, flowering and fruiting (Johnson & Mirza, 2020). Potassium is also needed in large quantities for metabolic processes such as photosynthesis, cell division, and movement of sugars and starches across different parts of plants (Zewide & Melash, 2021). Calcium is useful in the elongation and division of cells as well as formation of protein (Johnson & Mirza, 2020). Magnesium aids in production of adenosine triphosphate for production of energy and chlorophyll for absorption of light, which is important for plant growth (Johnson & Mirza, 2020). Lastly, sulphur is absorbed to carry out synthesis of proteins (Zewide & Melash, 2021).

Comerford (2005) describes the processes of nutrient transfer from its solid phase to eventual uptake by roots which forms part of a nutrient budget as illustrated in figure 2.1. Nutrient budgets illustrate the

necessity of balancing nutrients from their input to a crop ecosystem and eventual output for economic use. Any differences in inputs and output values results into a nutrient imbalance, which degrades the soil reducing its capacity to support plants (FAO, 2022). Determining nutrient availability in soil can be achieved through sampling and laboratory analysis and thereafter conducting spatial variability to model the trend of nutrients across a geographic region (Fernández & Hoeft, 2009).

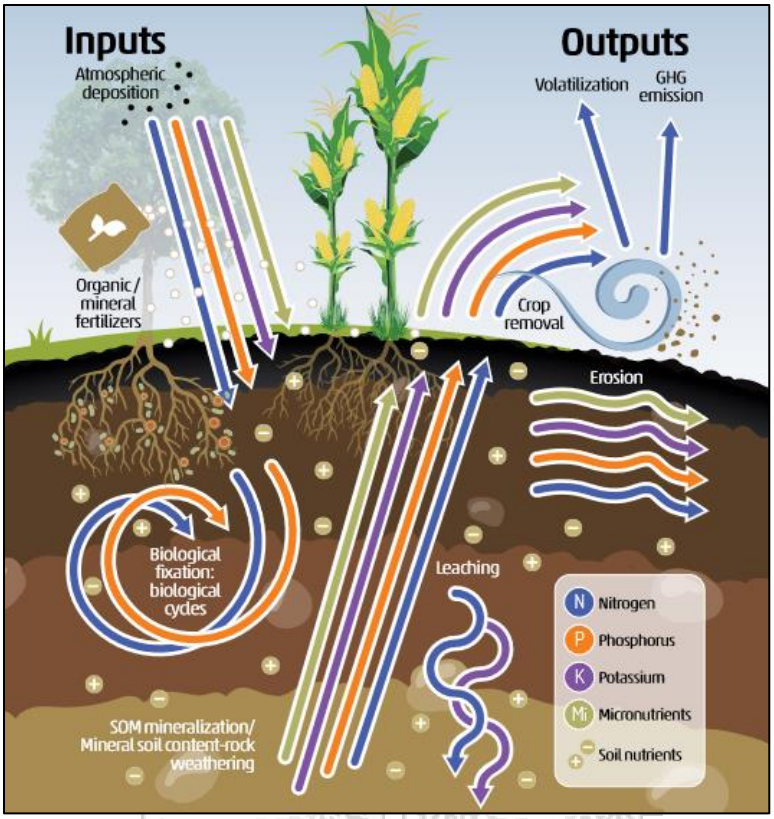


Figure 2.1: Nutrient Budget for a Crop Ecosystem (Source: FAO, 2022)

2.2.2. Plant Macronutrients for Human Nutrition: Case of Type-II Diabetes

Consumption of nutritious plant-based foods, produced from nutritious soil, improves the state of human nutrition. This is because of bioavailability of plant nutrients during digestion of these foods (Schönfeldt et al., 2016). There are limited attempts to extend the plant-human nutrition connection to include soil nutrition (Keesstra et al., 2016), due to various factors such as variation of food consumption patterns, complexity of the bioavailability process in human nutrition and variability of nutrient statuses in soils, plants and humans over time (Rekik & van Es, 2022). Berkhout et al. (2019) demonstrate this however, by modelling a relationship between soil properties, economic factors, and health conditions. It is determined that health condition (H) such as child stunting rates can be reduced by increasing amount of

soil nutrients (F) but can be influenced by indicators of economic development (C) of a region such as presence of a mining site or water body (Berkhout et al., 2019). This is represented in equation 2.1.

$$H = \alpha + \beta F + \gamma C + \varepsilon$$

Equation 2.1: Econometric Link between Soil Nutrients and Human Nutrition (Source: Berkhout et al., 2019)

Type-II diabetes is a non-communicable disease brought about by irreversible factors such as aging and genetics and reversible factors in lifestyle such as poor eating habits, smoking, and limited physical activity (Minari et al., 2023; Sami et al., 2017). In 2021, Kenya registered mortality of approximately 15,284 persons aged between 20-79 years to diabetes (Onteri et al., 2023; Sun et al., 2022). This can be attributed to prolonged diagnosis for diabetes and an increase in the reversible diabetes risk factors (Manyara et al., 2024). Vegetables have been identified as a rich source of micronutrients by both diabetic and non-diabetic individuals, however both groups report not consuming the required amounts to manage and reduce risk for diabetes (Kiprotich et al., 2022; Manyara et al., 2024). This is due to issues such as limited knowledge on recommended servings, unaffordability of healthy foods and food safety concerns (Manyara et al., 2024; Wekesah et al., 2019).

Increased intake of micronutrients has an influence on reducing the effects of non-communicable diseases (NCDs) (Nguyen et al., 2022). Table 2.1 presents the role of four soil macronutrients as identified in section 2.2.1 that have an influence on type-II diabetes, together with examples of plant foods that contain the nutrient.

Soil Macronutrient	Function	Plant Food Available
Potassium	Reduce rate of fat breakdown which may result to high blood sugar (Watson, 2022).	Beans, lentils, bananas, vegetables such as spinach, broccoli (Watson, 2022).
Calcium	Contains probiotics that reduces chances of resistance to secrete insulin (Jeon et al., 2018).	Spinach, kale, soybeans, white beans (Zelman, 2022).
Magnesium	Facilitates metabolism of carbohydrates which influences secretion of insulin (Dubey et al., 2020).	Nuts, whole grains, green leafy vegetables (Gray & Threlkeld, 2000).
Sodium	To regulate blood pressure levels with respect	Wheat (low-salt varieties of bread,

	to rate of potassium intake (Baqar et al., 2020).	pasta), rice, herbs e.g., basil and marjoram (Dansinger, 2023).

Table 2.1: Macronutrient Functions for Nutritional Management of Type-II Diabetes

2.3. Digital Soil Mapping

Digital soil mapping (DSM) involves the use of computational techniques to develop geographically referenced data on soil properties. It is a major improvement from conventional mapping procedures, facilitated by availability of geographical information systems (GISs), global positioning systems (GPSs), remote sensing techniques, predictive models, supporting software and geospatial soil data (Heung et al., 2021; Kienast-Brown et al., 2017). As illustrated in figure 2.2 and Kienast-Brown et al. (2017), the DSM process involves identifying an area for study, a set of collected soil samples and characteristics about the environment under study to serve as independent variables (covariates). The identified datasets are fed to an algorithm that selects the appropriate environmental features, and plots values of the soil property of interest to the environment, based on the selected covariate (Kienast-Brown et al., 2017). Sampled geographic locations are fitted with the respective soil property value, while unsampled locations will be predicted based on proximity to the sampled locations (Chagumaira et al., 2022). The output is a raster map, which is a grid of two-dimensional cells representing a unique geographical location by latitude and longitude (Heung et al., 2021). Each cell represents a value of the soil property under observation.

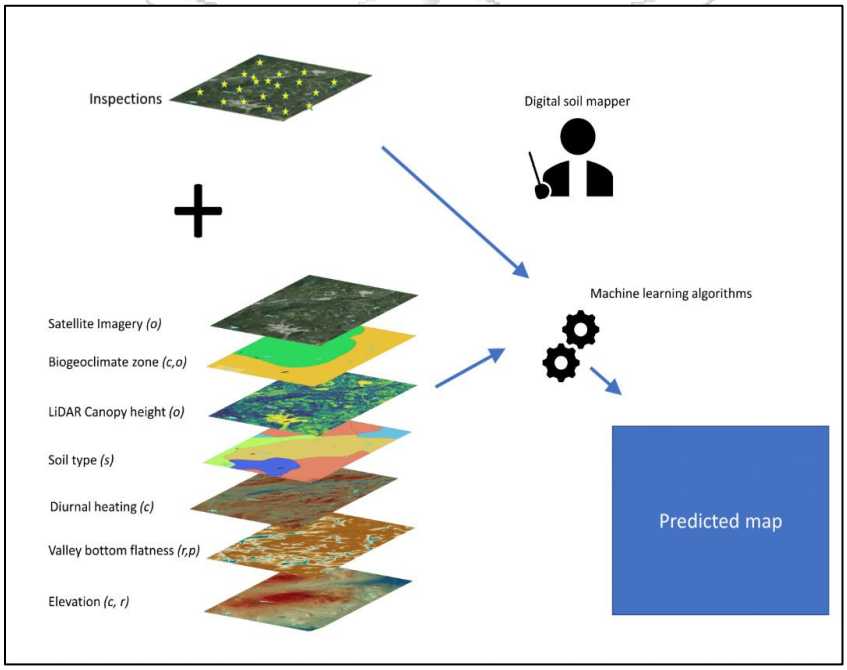


Figure 2.2: Process for Digital Soil Mapping (Source: Heung et al., 2021)

2.3.1. Input Data for Digital Soil Mapping

In spatial analysis of soil data, there are two sets of input. The first is of values of soil properties such as pH, organic carbon and nutrient components, which may take up the form of samples collected directly from the field and further analysed in a laboratory, or data that is already represented on geographical maps at a specified resolution (Hendriks et al., 2019). Legacy soil data takes in the form of the latter, where it features soil profiles that were collected and analysed using conventional surveying techniques. Use of this data as an approach to DSM is not as popular in the Kenyan context, since it might not provide an up-to-date view of soil properties, however Ngunjiri et al. (2019) and Minai et al. (2021) demonstrate its validity to achieve the same, especially when applying ML-based techniques for prediction. Hengl et al. (2017) can be considered a significant contributor to the field of legacy soil information within the Kenyan and Sub-Saharan context, by creating and availing a public repository of soil data and generating suitable soil maps that can be confidently applied in various related settings. This was made possible by the Africa Soil Profiles Database (AfSP) (Leenaars et al., 2014), a set consisting of 18,000 legacy soil profiles across Sub-Saharan Africa.

The second set of input is of covariates that describe the conditions of the environment under study, represented by the scorpan model in equation 2.2.

$$S_{c,a} = f\{s, c, o, r, p, a, n\}$$

Equation 2.2: Scorpan Model (A. B. McBratney et al., 2003)

The model is illustrated as classes (S_c) or attributes (S_a) of soil that are affected by:

i. Soil (s) – existing information on soils can be used in prediction of attributes and classes. This data can be derived from collected samples, ground based or remotely sensed spectral data and existing soil maps (Heung et al., 2021; Kienast-Brown et al., 2017).
ii. Climate (c) – this involves data layers such as average annual temperature and average annual precipitation that is collected from satellite-based sensors or climate models from weather stations (Heung et al., 2021; A. B. McBratney et al., 2003).
iii. Organisms (o) – this covariate factors for activities carried out by humans, plants, and animals. Data is captured via satellite and remote sensing imagery of vegetation and land cover (A. B. McBratney et al., 2003).Also, crop yield and forest inventory data can provide information such as soil moisture and aboveground biomass content (Heung et al., 2021).
iv. Relief (r) – this features details on the terrain of the site under study, such as wetness index, elevation, and slope length (Kienast-Brown et al., 2017). It is a popularly applied attribute (A. B.

McBratney et al., 2003), due to the accessibility of digital elevation models from digitised maps, ground measurements and remotely-sensed data (Heung et al., 2021).

v. Parent material (p) – this represents the parent material of the soil derived from sources such as digitised geological maps, measurement of natural gamma rays and earth fields such as gravitational and electromagnetic (A. B. McBratney et al., 2003).

vi. Age (a) – this attribute refers to the estimated time from which the soil developed from its source. It is rarely considered for DSM since it is difficult to suitably format this time for GIS, but it is suggested, in some cases, to model changes in human activity on the landscape over time and make inferences from them (Heung et al., 2021; Kienast-Brown et al., 2017).

vii. Spatial position (n) – considered a "valuable yet cheap source of environmental information" (A. B. McBratney et al., 2003), this attribute covers the proximity of coordinates as represented on a spatial map. It can be applied to generate values for unsampled locations based on proximity with sampled locations and estimate distances between a pixel to a geographical phenomenon to capture information based on the context of the landscape (Heung et al., 2021).

Covariate selection aims to enhance computational performance of the model (Lacoste et al., 2016), which enhances its reliability for replication in other settings (Xiaobo et al., 2010). Behrens et al. (2010) recommends the selection of few covariates to achieve acceptable prediction results. Pearson's correlation analysis is a common technique for selection, where the degree of linear correlation between two variables (Lu et al., 2019; Mishra et al., 2020). The formula is represented by equation 2.3, which demonstrates correlation between variables X and Y. S_{xy} denotes the covariance between X and Y while S_x and S_y represent their standard deviations (Kotu & Deshpande, 2019).

$$correlation(X,Y) = \frac{S_{xy}}{S_x\, X\, S_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x\, X\, S_y}$$

Equation 2.3: Pearson's Correlation Coefficient (Source : Kotu & Deshpande, 2019)

Output of the DSM process is presented as a raster map, which relies on the component of spatial resolution to represent area on the ground; for instance, a spatial resolution of 50m indicates that each cell represents a ground area of 50m by 50m (Heung et al., 2021). Use of high-resolution maps is encouraged due to the amount of information they contain enhancing accuracy of predicted maps (Buenemann et al., 2023). For example, distribution of soil organic carbon in farmlands of an administrative location in Kenya would be better represented in a higher resolution than a representation of the same for the country. However, this may present some pitfalls in analysis as highlighted in

Kienast-Brown et al. (2017), where the model may be prone to overfitting, increasing the time and storage used for processing. Selecting the optimal resolution can be achieved by determining the appropriate covariates of the environment under study, soil property being analysed, and geographical scope of the expected map (Piedallu et al., 2022).

2.3.2. Geostatistical Modelling Techniques used in Digital Soil Mapping

Geostatistical modelling in DSM primarily involves applying techniques for interpolation between sampled and unsampled locations, where values for the unsampled locations are generated based on proximity to the sampled points (Kienast-Brown et al., 2017). They are stochastic in nature; in that they rely on statistical information of sampled data points to determine the extent for interpolation (Brindha et al., 2023). This varies from deterministic models that apply a weighting technique to unsampled points, where those nearer to the sample are given higher weightage and decrease as they move farther from the sampled location (Sahu et al., 2021). Classic techniques for DSM by geostatistical modelling is ordinary kriging (OK) and deterministic modelling is inverse distance weighting (IDW), and several studies on mapping soil nutrients implement both since they can be used interchangeably (Barrena-González et al., 2022; Ouabo et al., 2020; Sahu et al., 2021).

IDW applies a distance reverse function, which assumes that distance between neighbours is proportional to similarities and rate of correlations between them (Yasrebi et al., 2009). It is more suited for evenly distributed points, and it is possible to influence the rate of interpolation by enhancing weighting power of significant points on the map. Sampled points with greater weighting powers will not influence each other but will affect unmeasured points which have smaller powers generated based on distance to measured points (Ouabo et al., 2020). Equation 2.4 describes IDW where u is the estimation location, u_i are the locations of sampled points within the search neighbourhood that has n sample points, $Z^*(u)$ is the inverse distance estimate at u, $Z(u_i)$ is conditioning data at sample points, and λ_i are weights assigned to each sample point (Babak & Deutsch, 2009).

$$Z*(u) = \sum_{i=1}^{n} \lambda_i Z(u_i)$$

Equation 2.4: Calculation for Inverse Distance Weighting (Source: Babak & Deutsch, 2009)

The weights are determined as in equation 2.5, where d_i represents Euclidean distances between u and u_i and exponent p is 'rate' of decreasing weight whose value ranges are recommended to be between 1 and 4.Sum of all inverse weights λ_i is equal to 1, and the accuracy for interpolation is affected by the power

parameter *p,* size of search neighbourhood and number of neighbours (Babak & Deutsch, 2009; Ouabo et al., 2020).

$$\lambda_i = \frac{\left(d_i^p\right)^{-1}}{\sum_{i=1}^{n}\left(d_i^p\right)^{-1}}, (i, \dots, n)$$

Equation 2.5: Inverse Weight Calculation Function (Source: Babak & Deutsch, 2009)

OK, conversely, applies a similar process of interpolation to IDW, the difference comes in the weighting technique, where it applies the spatial information of measured points to create relationships between them. This relationship compares the distance between points (separation distance) and variance in value, and is compared as semivariance $\lambda(x_i, x_j)$ where x_i and x_j are two measured data points (Kumar & Sinha, 2018; Ouabo et al., 2020). Semivariance is denoted in equation 2.6, where x represents the sampling locations and z(x) is the observed value.

$$\lambda(x_i, x_j) = \frac{\left[z(x_i) - z(x_j)\right]^2}{2}$$

Equation 2.6: Semivariance Function (Source: Kumar & Sinha, 2018)

The semivariance is then related to the separation distances throughout the entire set of measured values. This is applied through a semivariogram model, as in equation 2.7, where average semivariance for a separation distance h is derived from n pairs of samples that have the same distance h (Kumar & Sinha, 2018).

$$\hat{\lambda}(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [z(x_i) - z(x_{i+h})]^2$$

Equation 2.7: Semivariogram Model for Ordinary Kriging (Source: Kumar & Sinha, 2018)

2.3.3. Machine Learning Modelling Techniques used in Digital Soil Mapping

Machine learning (ML) has been largely adopted in DSM largely due to their ability to outperform geostatistics and other classical statistical methods. Not only can they handle complex nonlinear relationships between soil covariates but inter-related environmental covariates, but can also model data without much statistical assumptions on the distribution, making them effective for modelling within small-scale geographic extents and using legacy data (S. Chen et al., 2022; Wadoux et al., 2020). Plus, they utilise current trends in computing such as cloud computing and parallel computing to generate soil maps from big data (S. Chen et al., 2022). Padarian et al. (2020), however, points out that the use of

large datasets for modelling by ML may affect accuracy of prediction, creating the need to select appropriate covariates to improve how generated maps are interpreted.

Support vector machine (SVM) is a popular ML technique for DSM, due to its effectiveness with small to medium datasets and high-dimensional data, making it less prone to overfitting (Folorunso et al., 2023; Lamichhane et al., 2019). Depicted in figure 2.3, it is a classification and regression technique where a decision surface (hyperplane) is created to separate data items based on classes. Among the surfaces generated those that create a large margin between classes is regarded as optimal, while those towards the edges of classified data points are termed as support vectors (Li et al., 2014; Nikparvar & Thill, 2021). This, however, may not be as straightforward, due to outliers in the dataset, therefore misclassification is allowed while penalising them based on their distance from the margins (Nikparvar & Thill, 2021). Regardless, high generalisation of the algorithm is achieved, since all data points including unseen data is mapped (Khaledian & Miller, 2020).

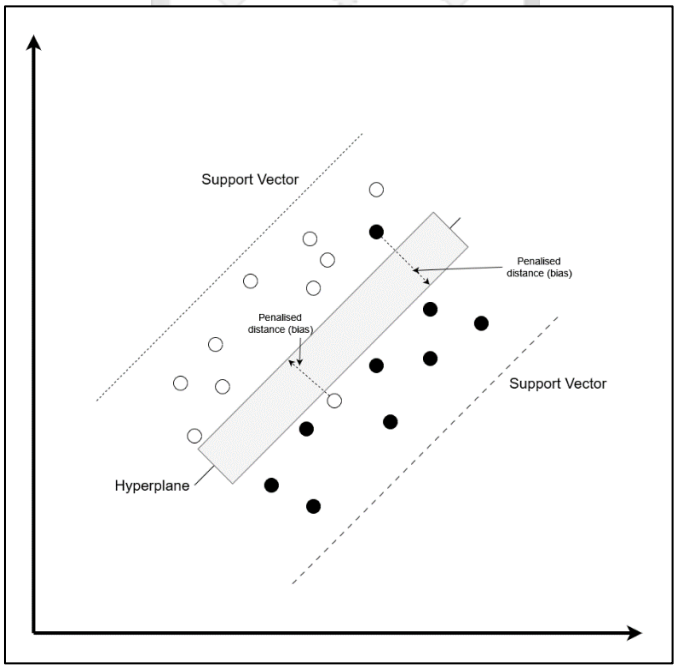


Figure 2.3: Process of Support Vector Machine Classification (Adapted from Li et al., 2014 and Nikparvar & Thill, 2021)

The regression feature of SVM is more suited for DSM (Khaledian & Miller, 2020), and is illustrated in figure 2.4 where $K(X, X_m)$ represents the kernel whose function is to translate vector points (X, X_m) to a higher-dimensional plane (Y). Its objective function is to carry out this translation while achieving the widest margin possible and minimising the bias(b) arising from allowed misclassification of outliers (Nikparvar & Thill, 2021). Lamichhane et al. (2019) indicates that selecting an appropriate kernel as well as hyperparameters for tuning can be computationally expensive. In DSM the commonly used

kernel is a radial basis function together with bandwidth (γ) and penalty (C) as hyperparameters (Emadi et al., 2020; Suleymanov et al., 2021).

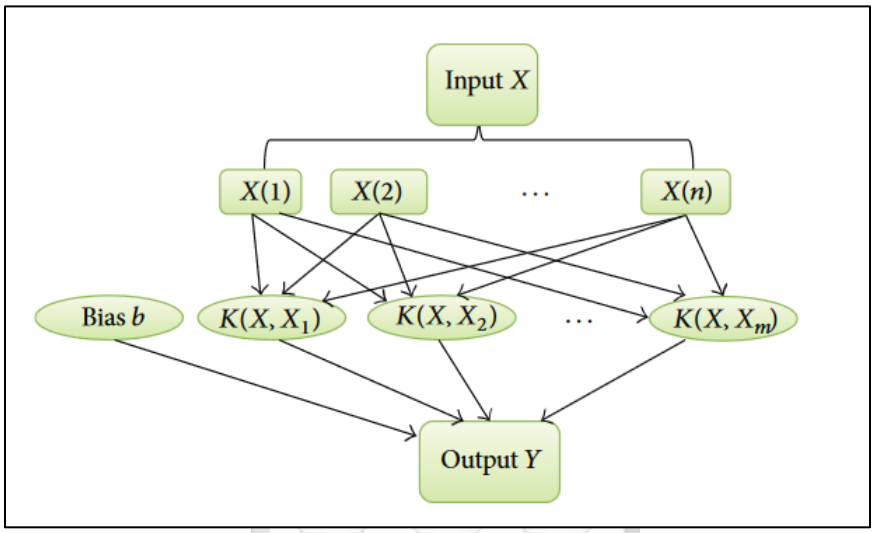


Figure 2.4: Structure for Support Vector Machine Regression (Source: Li et al., 2014)

2.3.4. Hybrid Modelling Techniques used in Digital Soil Mapping

Despite the advantages of ML techniques over geostatistical methods in prediction accuracy, Chen et al. (2022) highlights how geostatistical methods can still be beneficial in capturing information on spatial structures. This information can easily be overlooked by use of a single ML approach in the case working with highly variable data, resulting in erroneous predictions (Tarasov et al., 2018). Lamichhane et al. (2019) reports the popularity of regression kriging (RK) as a hybrid technique in DSM. It features a regression model, such as multiple linear regression (MLR), to assess the deterministic trend between soil property (target attribute) and environmental covariates (independent variables), and a kriging function to interpolate residual values that are stochastic in nature (FAO, 2018). This is presented in equation 2.8 where γ represents the target attribute, X_i denotes the environmental covariates as predictors, β_i are coefficients that represent relationships between the target and independent variables, and ε signifies residual values resulting from errors in prediction (FAO, 2018).

$$\gamma = \sum_{i=0}^{n} \beta_i X_i + \varepsilon$$

Equation 2.8: Regression Kriging (Source: FAO, 2018)

RK is interchangeably referred to as multiple linear regression kriging (MLRK) to capture a set of numerous predictors and coefficients being applied in the model (Bolat et al., 2020). The mapping process by RK is represented in figure 2.5, where a map of soil property values is first layered over

identified environmental covariates and prepared to a tabular format for analysis by the regression model. Suitable covariates are selected based on strength of collinearity between variables. Any errors derived from prediction are finally added to the model output after a kriging process (FAO, 2018).

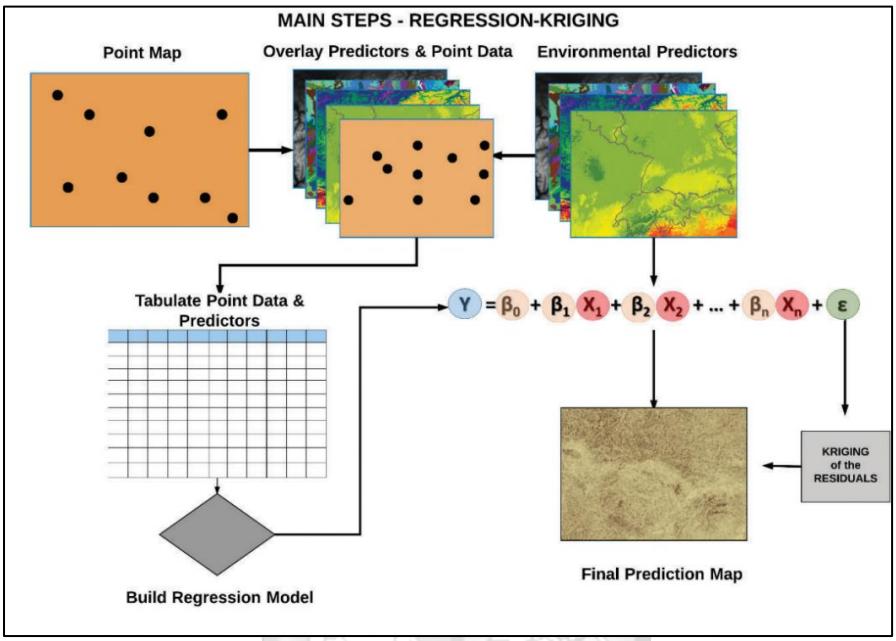


Figure 2.5: Process for Regression Kriging (Source: FAO, 2018)

2.3.5. Performance Measurement of Predictive Models for Digital Soil Mapping

Evaluation of geospatial models is largely done using the root mean squared error (RMSE) (S. Chen et al., 2022), as represented in equations 2.9. O_i and P_i represent observed and predicted values of soil properties while O_{avg} and P_{avg} represent their respective average values and n is the number of samples (Suleymanov et al., 2021). In evaluation, the most accurate models considered for prediction are those that return RMSE values closer to 0 (Gao et al., 2021).

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n}(O_i - P_i)^2}{n}}$$

Equation 2.9: Root Mean Squared Error (Source: Suleymanov et al., 2021)

2.4. Summary of Related Works and Conceptual Framework

To guide the study, several works were identified that closely relate to determination of food items and use of hybrid geospatial models in prediction and mapping of soil nutrients. They are illustrated in table 2.2, with the highlights indicated from each study that informed the development of the research.

Study	Techniques	Results	Highlights
(Karega, 2022)	Dataset: 651 food items Low rank matrix factorization	100.0% prediction accuracy	Need to enhance food predictive tool to utilise location-specific data in generating food items.
(Suleymanov et al., 2021)	Dataset: 76 collected topsoil samples Covariates: Relief –elevation, slope, profile curvature Models: SVM, Multiple Linear Regression (MLR), Ordinary Kriging (OK)	Calcium (Ca): 5.84 MLR v 5.33 SVM RMSE Magnesium (Mg): 10.05 MLR v 2.09 SVM RMSE	ML techniques (SVM) can be applied in spatial variability and generate similar maps to those by geostatistical techniques (OK). Individual use of ML techniques is likely to not cater for spatial structures of observed data (Tarasov et al., 2018).
(Gao et al., 2021)	Dataset: 2538 collected soil samples Models: OK, MLR, multiple linear regression kriging (MLRK) Covariates: Relief – elevation, slope, topographic wetness index; Climate – mean average precipitation, mean average temperature; Organisms – normalised differential vegetation index, vegetation curvature	Available potassium (AK): 33.9677 for OK RMSE, 43.9463 for MLR RMSE, 34.1836 for MLRK RMSE	Hybrid machine learning and geostatistical techniques generate more accurate maps than single ML or geostatistical techniques. OK can generate reliable results in cases of limited costs for modelling.
(Minai et al., 2021)	Dataset: 76 legacy soil profiles Covariates: Relief - Slope, Soil - Landsat Band7	Soil organic carbon (SOC): 0.43 for SMLR RMSE, 0.02	Legacy soil data can be applied to generate soil maps. OK outperformed other

	Models: stepwise multiple linear regression (SMLR), OK	for OK RMSE	models by failed to capture spatial variability of soil properties.
(Patriche et al., 2023)	Dataset: 1444 samples from LUCAS 2015 soil database Covariates: Relief – elevation, slope, topographic wetness index; Organisms e.g. normalised differential vegetation index. Models: OK, MLRK	Available phosphorous (P): 19.323 for OK RMSE, 18.583 for MLRK RMSE Extractable potassium (K): 193.815 for OK RMSE, 199.578 for MLRK RMSE	MLRK and OK based models can produce highly similar results.

Table 2.2: Summary of Related Works

In this study, a MLRK model was developed to create georeferenced maps of identified soil macronutrients. Input for the model will be soil macronutrient information and location data captured as georeferenced soil profiles. The model was tested against three other models based on OK, SVM and MLR, to produce maps that depict soil nutrient information. Accuracy of the developed models was tested using the RMSE index as identified in section 2.3.5 and compared against each other to determine the most accurate in generating soil maps. Lastly, a tool for determining suitable vegetables based on their macronutrient content was developed, given soil macronutrients and location information. Vegetable nutrient information was retrieved from an external source of identified food items consumed in Kenya, while soil nutrient information was from the generated soil nutrient maps. Location information was input as coordinates from the user and passed to the maps to derive nutritional information based on a specified location. The expected output is of plant vegetables based on a complement of soil and food nutrients. This workflow is depicted in figure 2.6.

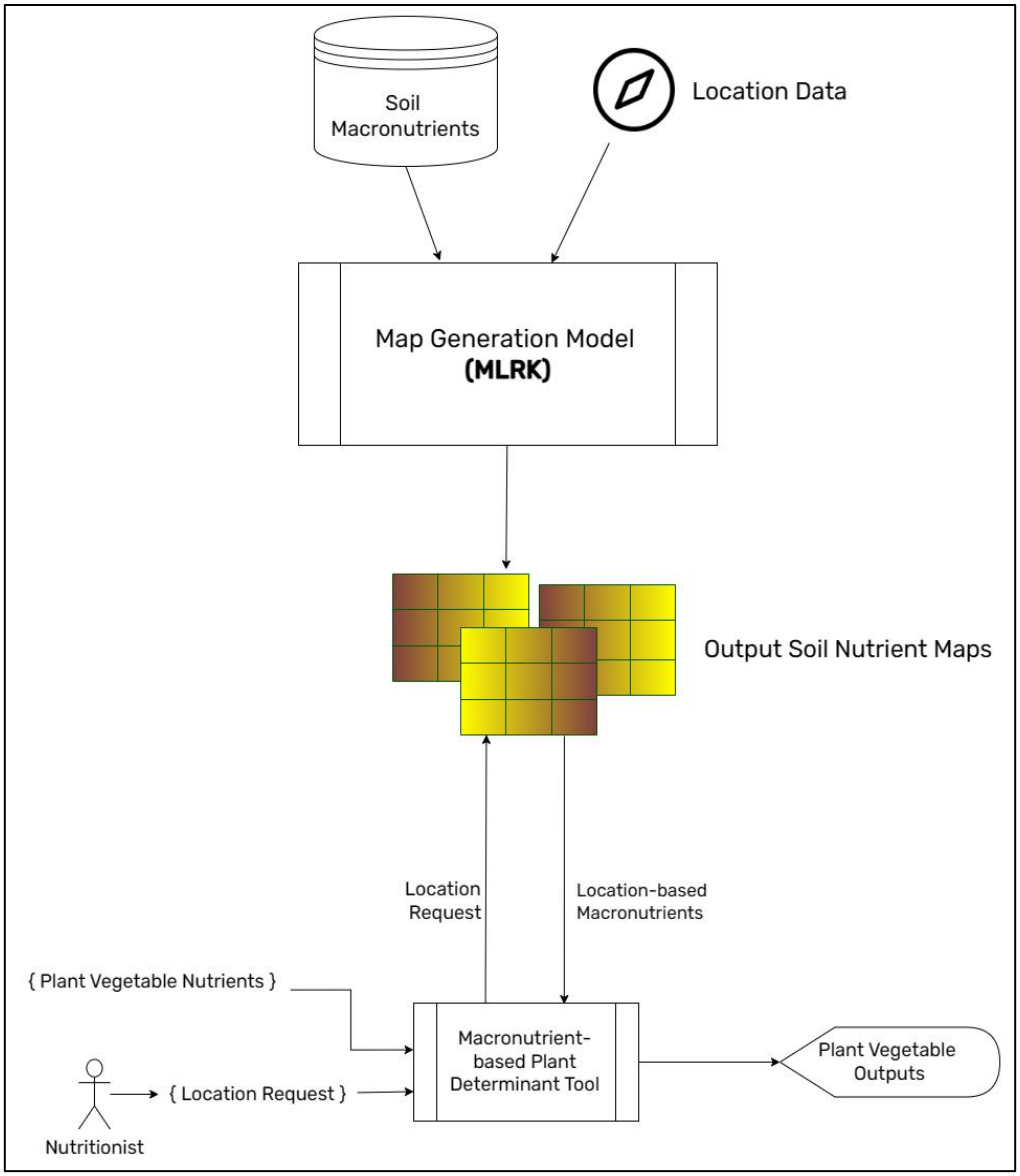


Figure 2.6: Conceptual Framework for the Study

Chapter 3: Methodology

3.1. Introduction

This chapter details the approach for carrying out the study to achieve the fourth and fifth objectives as indicated in section 1.3.2. It describes the process for conducting the study through the research design, the identified population and sampling strategy and methodology for developing the system. Concerns around ensuring quality of the research including standards for maintaining ethical conduct are addressed.

3.2. Research Design

The study applied an experimental research design, which looks at how the response of a system is influenced by varying related factors. This relationship was evaluated by use of statistical techniques to determine the most influential variable(s) that lead to an optimal desired output from the system under test (Hanrahan et al., 2005). As described in Tedre & Moisseinen (2014) experiments occurred in three forms throughout the project as controlled, comparison and trial experiments. The controlled experiment will be featured in the selection of appropriate environmental covariates for the geospatial models, which affect the accuracy of predicted soil maps. A comparison experiment was conducted in testing for prediction accuracy of the hybrid geospatial model against other developed models (OK, MLR and SVM-based). These two experiments were instrumental in answering the fourth research question as in section 1.4. Lastly, a trial experiment was conducted in the final phase of testing the validity of the developed geospatial models. Maps developed from the model were used to provide soil nutrient levels based on geographical location. This information on nutrients were then applied to determine suitable vegetables depending on their nutritional contents.

3.3. Population and Sampling

The identified population is samples of Kenyan soil and food items consumed in the country. Soil samples and their covariates are recorded as georeferenced soil profiles and satellite images. The soil nutrient maps identified for this study are those for Sub-Saharan Africa, where maps of exchangeable magnesium and potassium were applied (Leenaars et al., 2014). Sources of covariates such as elevation, forest cover and climate are captured as images at 250m resolution were selected depending on their relevance in prediction to the soil nutrient maps (Hengl, Leenaars, et al., 2017). The data source

identified for food items is the Kenya Food Composition Tables dataset, which describes food items using their components, such as mineral and vitamin content, categorised according to their food category e.g., tubers, cereals and meats (FAO & Government of Kenya, 2018).

The mode of sampling selected for this study was stratified random sampling which classified the population into specified categories based on a criterion defined by the researcher. Population on the soil samples was first categorised by location while those by food items will be classified by the food group, which was identified as vegetables. Random sampling from each group ensured adequate sampling across the population and allowed for estimation to be done within and between population classes (Howell et al., 2020). This greatly improved the final accuracy of the resulting output.

3.4. System Development Methodology

The development methodology applied was this study is Cross Industry Standard Process for Data Mining (CRISP-DM), which is a standard and industry-independent process model for data science related projects (Schröer et al., 2021). As illustrated in figure 3.1, the lifecycle model consists of six iterative phases, that are 'loosely connected' to allow for flexibility in project implementation (Saltz, 2021).

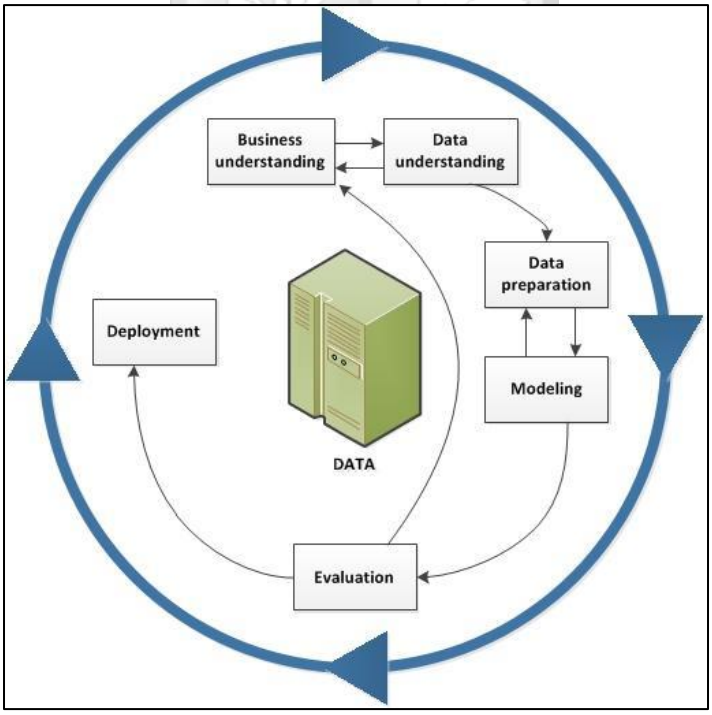


Figure 3.1: CRISP-DM Process Model (Source: IBM Corporation, 2021)

The activities for each phase with relation to the project are outlined as follows (Saltz, 2021; Schröer et al., 2021):

i. Business understanding – this involved identifying the goal and requirements of the project and how it relates to achieving business objectives. For this study, the desired outputs were a georeferenced maps of soil macronutrients that were then be applied to develop a tool by which nutritionists can determine plant vegetables based on location and nutritional requirements.

ii. Data understanding – this phase focused on gathering data relevant to the project, inspecting its quality, describing its properties and attributes, as well as establishing relationships between attributes. In the study, the identified population in section 3.3 was further explored in and descriptive statistics generated to get a better understanding of how the data was utilised throughout the project.

iii. Data preparation – here, the collated datasets were appropriately selected, cleaned, and formatted to enhance its quality for modelling. For this research, the soil databases as indicated in section 3.3 were cleaned to feature relevant data points on nutrients based on a set of environmental covariates that will be selected in the data understanding phase. Also, the database on food composition was prepared to only include relevant information on food items and the respective nutrient values.

iv. Modelling – this stage involved selecting techniques for analysing the prepared data and determining the approach for analysis i.e., developing datasets for training and testing the model. Thereafter, the analysis models were built, and their performances assessed based on a set of criteria. This workflow will be applied in this study in two rounds, the first to handle modelling of geospatial data to generate digital maps of soil nutrients, and the second to cater for determination of food items based on location and nutrient data.

v. Evaluation – at this phase, the results were checked against the project plan, goals, and requirements to verify their alignment with the objectives. Within the study, this was done by testing the developed models for accuracy in predicting soil nutrient values and determining food items.

vi. Deployment – this is the last stage that involved releasing the developed models for public use. When the study matured to this phase, a final review was conducted on the development of these models and potential areas for improvement in future projects were highlighted. Before public

release, the developed tools will go through scheduled phases of monitoring and maintenance to ensure consistency and reliability in their operation.

3.5. Research Quality

The study built on existing techniques for data modelling and analytics, which guaranteed generation of valid output. Models developed were tested for accuracy until they achieved an acceptable level of generating the required output. This was done using identified metrics for evaluating accuracy and a set of test data. The research was also documented to ensure replication in other related studies.

3.6. Ethical Considerations

This study utilised open-source datasets on spatial soil maps and food nutritional information, as identified in section 3.3, guided by the conditions set by copyright. Before conducting the study, it was subjected to ethical approval by the Strathmore University - Institutional Scientific Ethics Review Committee (SU-ISERC) to ensure alignment with Strathmore University's research policies as well as those by the National Commission for Science, Technology, and Innovation (NACOSTI). This process was acknowledged as demonstrated in appendix A. Previous works that were used to guide this research were acknowledged through citation of content and media. The originality report generated for this study is illustrated in appendix B.

VT OMNES VNVM SINT

Chapter 4: System Analysis and Design

4.1. Introduction

This chapter describes the process and results of analysis to generate requirements that guided development of the system. It also includes conceptual designs of the system structure, its components and operations illustrated through Unified Modified Language (UML) diagrams. It also covers wireframes that were used to guide the development of the food determinant tool.

4.2. System Requirements Analysis

The goal of the study was to use georeferenced maps of soil macronutrients in developing a predictive tool that determines vegetable derived macronutrients for a diabetic patient. The process for requirements was primarily conducted by document analysis, which involved extensive review of existing literature regarding spatial analysis of soil nutrients and its application in prediction of vegetables for diabetic patients. Two categories of requirements were evolved from this process: functional requirements, which focus on the operations of the system to support a user in completing a task, and non-functional requirements that cover properties the system should possess to facilitate the base system functionalities (Dennis et al., 2012).

4.2.1. Functional Requirements

i. The geospatial model should predict soil nutrient values at unseen locations from existing soil samples.
ii. The food classifier should categorise vegetables by nutrient value.
iii. The application should allow users to set a location.
iv. The application should retrieve nutrient values based on the location set by the user.
v. The application should determine suitable vegetables based on soil nutrient value.

4.2.2. Non-functional Requirements

i. Compatibility: The application should be responsive to various platforms i.e. desktop and mobile.
ii. Performance: The application should quickly respond to user queries on food predictions.
iii. Usability: The application should allow for quick and easy interactions via its interface.

iv. Reliability: The models supporting the application should provide consistent and accurate results at each instance of prediction.

v. Availability: The application interface and supporting logic for operation should experience minimal downtime.

4.3. System Design

Developing designs for the system took up an object-oriented approach, where system entities are viewed as objects which interact with each other. Similar objects are organised into classes, which have attributes (data values) and methods (operations) that guide the interaction of objects when instantiated (Gomaa, 2011). The benefits of applying this approach is the ability to reuse and modify objects without much impact on other objects, which is suitable for agile processes in system design and maintenance (Kendall & Kendall, 2011; Sommerville, 2011). Diagrams resulting from the design process are the system architecture, use case, class, sequence, and activity diagrams as well as the database schema.

4.3.1. System Architecture

Figure 4.1 presents a layered architecture for the developed system. It comprises of layers A to D each with identified roles within the overall system. Layer A consists of core support components i.e. the two primary databases for generating soil maps as well as prediction of foods. Layer B contains the business logic for the application, which features the two models for generating soil maps and determination of foods. Layer C and D are user-centric, with C containing components to handle user requests and responses, and D hosting the interface which they interact with. This layered architecture is beneficial in allowing for component independence, where modifications done in one layer are less likely to affect operations in another layer. However, since they operate only through the base repositories in the core support components, there will be a need to enable distributed processes when scaling the system to avoid a single point of failure (Sommerville, 2011).

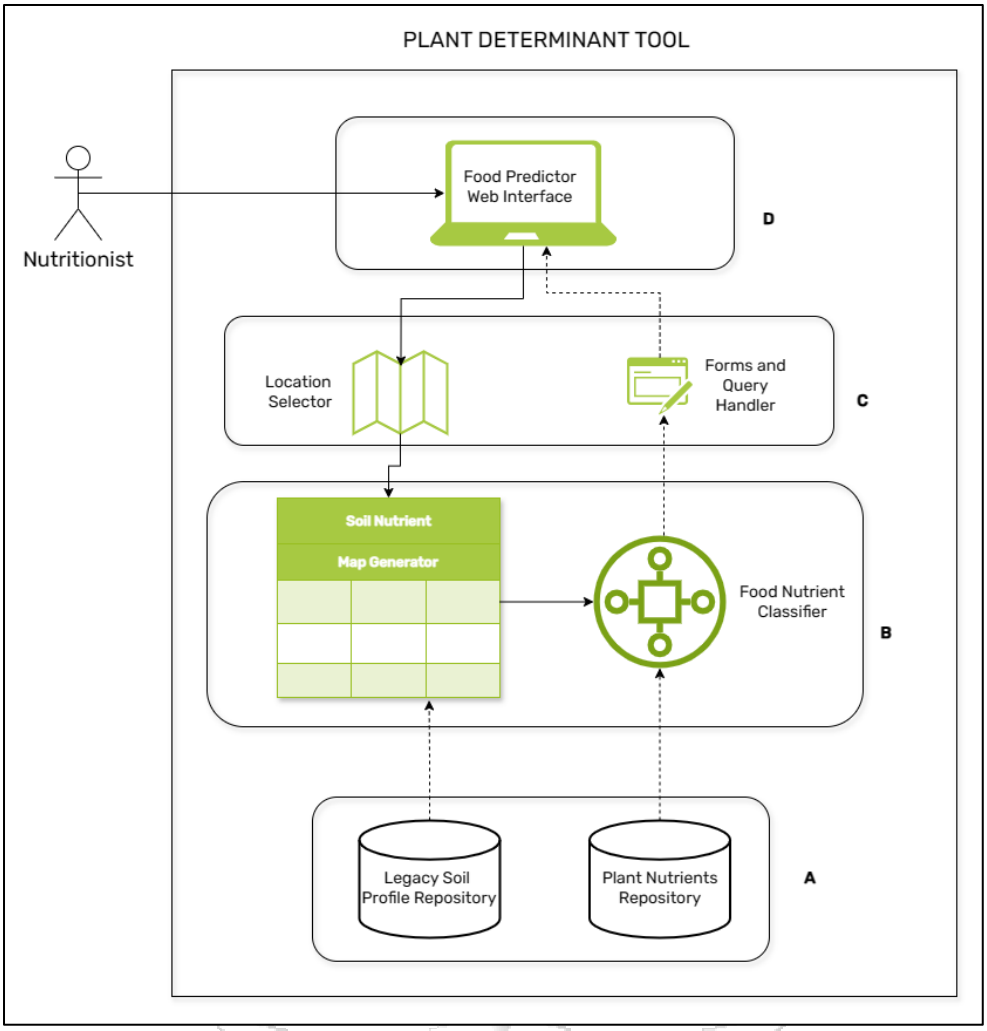


Figure 4.1: Plant Determinant Tool System Architecture

4.3.2. Use Case Diagram

Figure 4.2 describes the use cases for the developed system, comprising of all actors within the system and the sequence of interactions carried out by each of them. The system largely features one human actor, the nutritionist who provides location data for the patient and receives suitable vegetables based on the soil macronutrient and plant macronutrient values. The other two actors are machine-learning based models – the soil map generator that predicts soil nutrients and returns results as soil maps, and a nutrient classifier for predicting foods based on their nutrient values comparing to generated soil nutrient values. Detailed descriptions of a use case from each actor is illustrated in tables 4.1, 4.2 and 4.3.

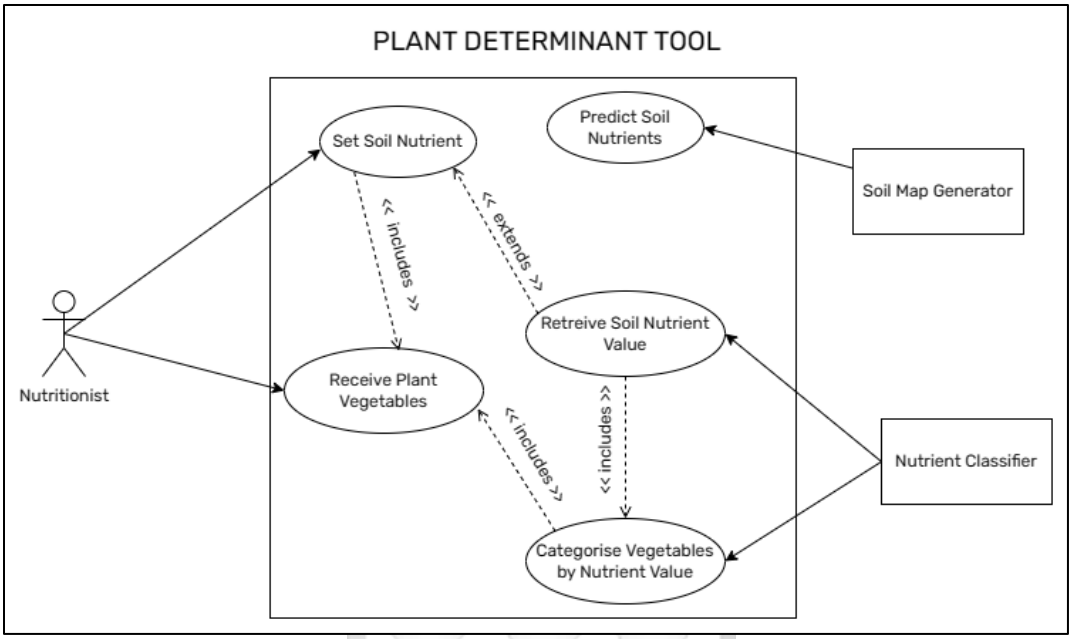


Figure 4.2: Plant Determinant Tool Use Case Diagram

Use Case Name	Categorise Vegetables by Nutrient Value
Actor	Nutrient Classifier
Dependency	Select Soil Nutrient Value
Precondition	Nutrient Classifier has identified a soil nutrient value.
Main Sequence	1. Classifier retrieves set of foods categorised by nutrient values. 2. Classifier compares soil nutrient to nutrients in each category. 3. Classifier provides category matching soil nutrient value.
Postcondition	Classifier generates list of foods in matched category.

Table 4.1: Categorise Vegetables by Nutrient Value Use Case Description

Use Case Name	Receive Plant Vegetables
Actor	Nutritionist
Dependency	Set Soil Location, Categorise Vegetables by Nutrient Value
Precondition	Nutritionist has provided a location and classifier has generated list of foods in matched category.
Main Sequence	1. Nutrient classifier organises set of predicted foods by nutrient value.

	2. Nutrient classifier filters the first ten food items. 3. Nutrient classifier sets food items to table on interface. 4. User directs to table of predicted foods.
Postcondition	None

Table 4.2: Receive Food Recommendations Use Case Description

Use Case Name	Predict Soil Nutrients
Actor	Soil Map Generator
Dependency	None
Precondition	None
Main Sequence	1. Generator selects soil profiles from database. 2. Generator selects environmental covariates. 3. Generator carries out prediction of soil nutrients at unsampled locations.
Postcondition	Map generator saves predicted maps.

Table 4.3: Predict Soil Nutrients Use Case Description

4.3.3. Activity Diagram

An activity diagram describes the flow of activities in a specific use case based on the use case diagram. Figure 4.3 demonstrates the process for the set soil nutrient process done by the nutritionist. The user selects coordinates from a map and is then provided with two options: to generate the potassium value from the soil potassium map or the magnesium content from the soil magnesium map. The soil nutrient content is then retrieved and passed on to determine suitable food items.

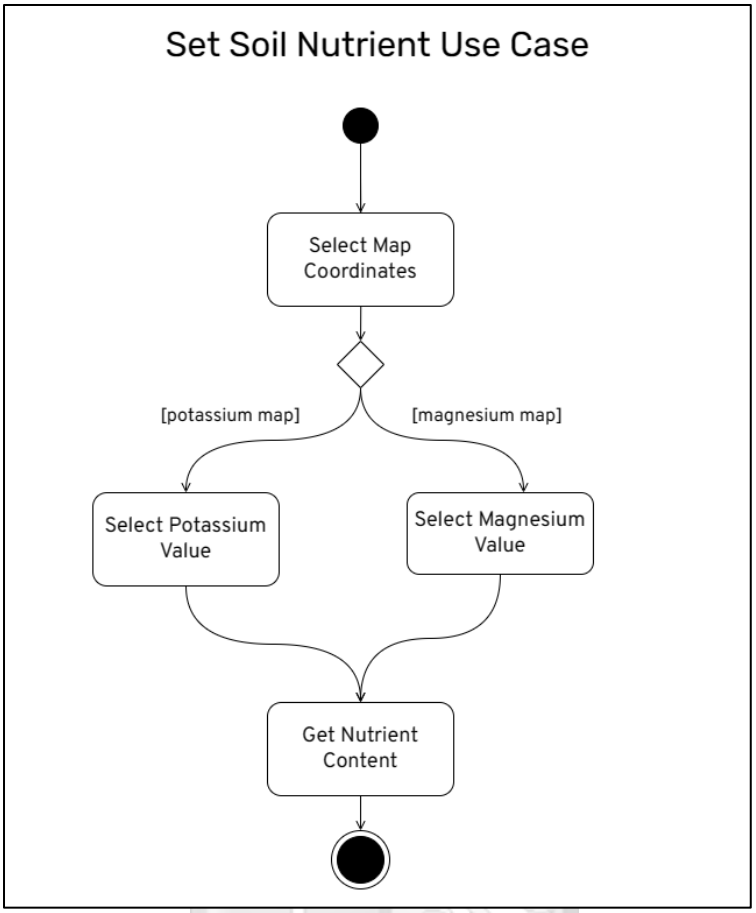


Figure 4.3: Activity Diagram for Set Soil Location Use Case

4.3.4. Sequence Diagram

A sequence diagram illustrates a set of interactions between classes or object instances over time and is derived from use case analysis to demonstrate the processing as described (Kendall & Kendall, 2011). Figure 4.4 depicts the sequence diagram for the receive food recommendation use case. The nutritionist triggers the process by providing a request containing location information. This is then handled by the soil map service to retrieve the nutrient content at the specified location. Nutrient values are then passed to the food item service to generate the list of predicted foods that is eventually output to the nutritionist.

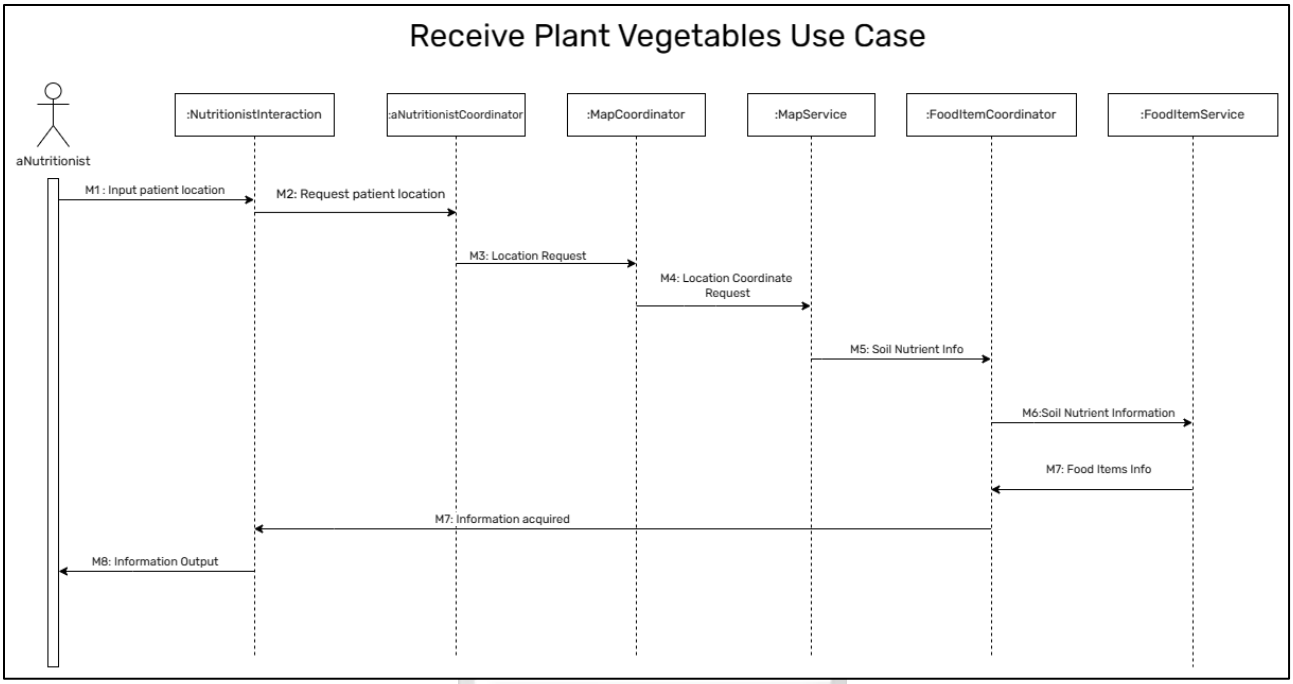


Figure 4.4: Sequence Diagram for Receive Food Recommendation Use Case

4.3.5. Class Diagram

Figure 4.5 describes the four classes for the developed system, which contain of individual attributes and methods to carry out data operations. It also features the associations between classes. The soil profile class consisting of recorded soil information at a particular location is used to generate soil maps that are consequently used in providing food recommendations based on nutrient value. The recommendations class is also supported by the foods class as it contains data on recommended food items.

VT OMNES VNVM SINT

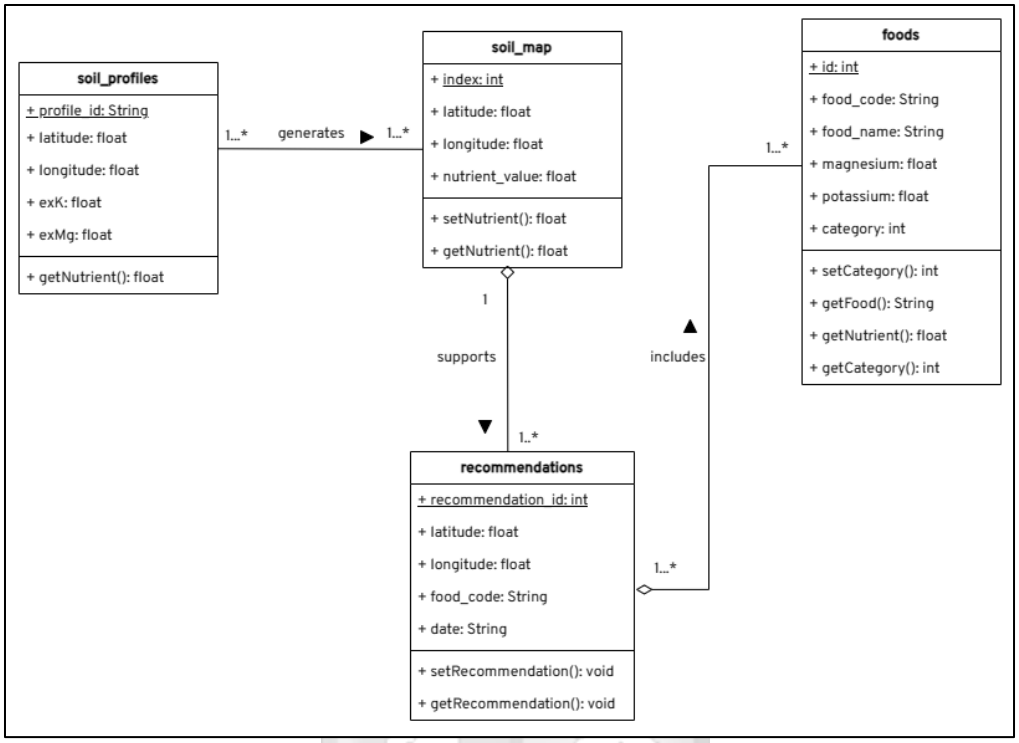


Figure 4.5: Plant Vegetable Determinant Tool Class Diagram

4.3.6. Database Schema

A database schema is a logical representation of a system's data model. Derived from the class diagram, it represents the classes as relations with the attributes as columns and individual data objects are stored as rows. This is illustrated in figure 4.6, where data types for items stored in each column are indicated. Relationships between relations are also represented as those in the class diagram, for instance only one soil map can be used to generate at least one food recommendations.

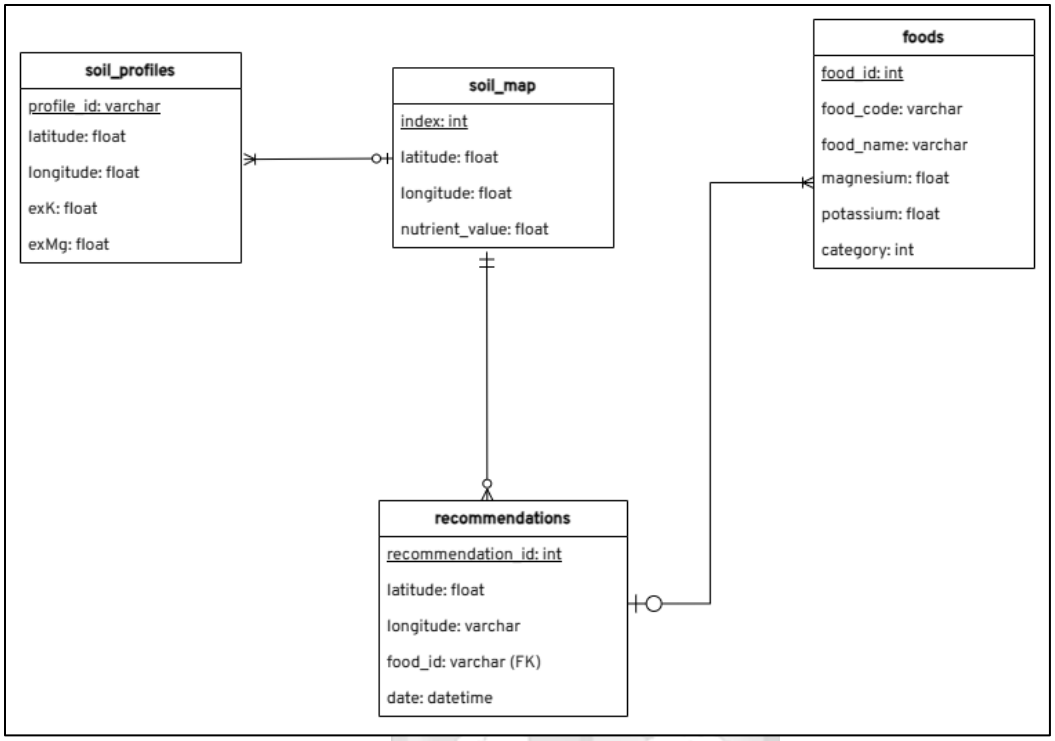


Figure 4.6: Vegetable Plant Determinant Tool Database Schema

4.3.7. System Wireframes

Figures 4.7 to 4.10 depict designs that were used to develop interfaces that the nutritionist interacts with as they use the tool. This was based on their use cases as in figure 4.2. Figure 4.7 demonstrates the location selector interface where they will input coordinates based on a set geographic location. There is a boundary of a specified region that has been indicated to specify the region in which the soil macronutrients have been analysed.

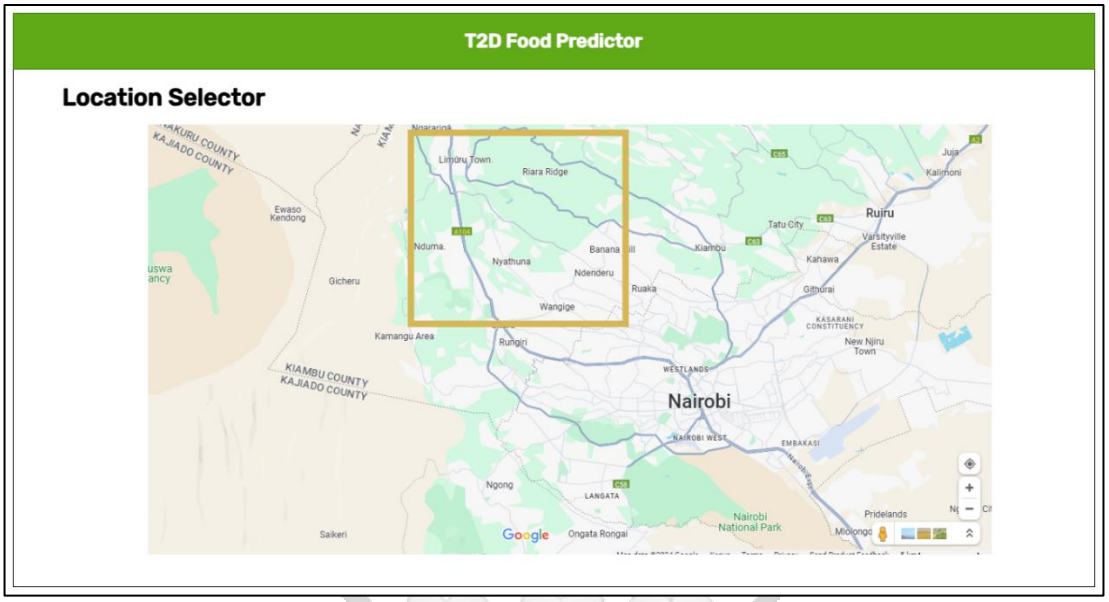


Figure 4.7: Location Selector for Vegetable Plant Determinant Tool

After selecting the nutrients, they are directed to the Food Generator form as in figure 4.8, where they are prompted to generate either potassium or magnesium nutrient content. This is then passed on to the nutrient classifier model to categorise suitable vegetables that meet the nutrient values.

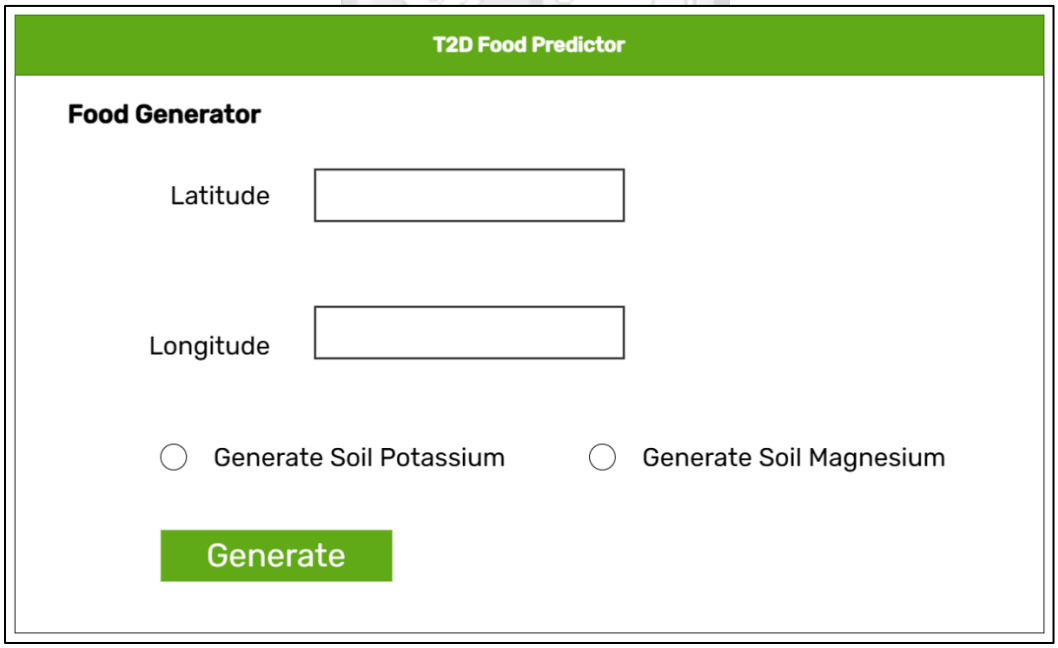


Figure 4.8: Food Generator Form

After the classifier determines suitable vegetables, they are then presented to the 'Predicted Vegetables' table as illustrated in figure 4.9. Here, the nutritionist user can select the suitable foods to recommend to the patient.

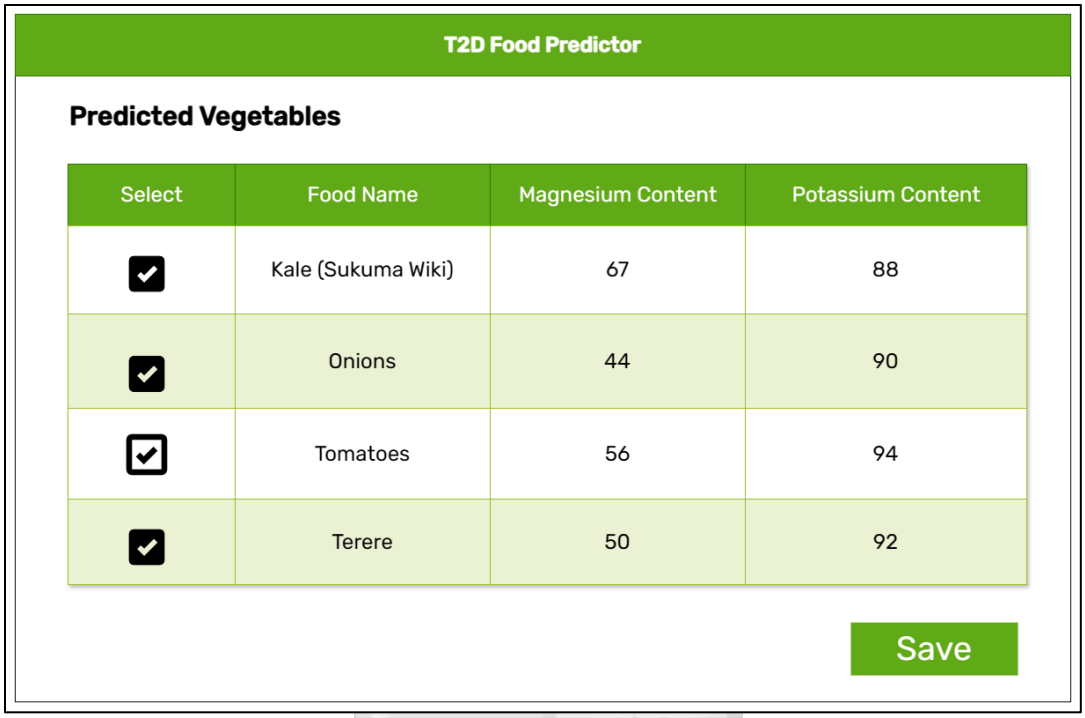


Figure 4.9: Predicted Vegetables Table in Food Determinant Tool

All recommendations made by the nutritionist can also be viewed in the 'Previous Recommendations' table as in figure 4.10. This table includes the date of recommendation, coordinates of the location at which the vegetables were generated, and a list of foods selected by the nutritionist after prediction.

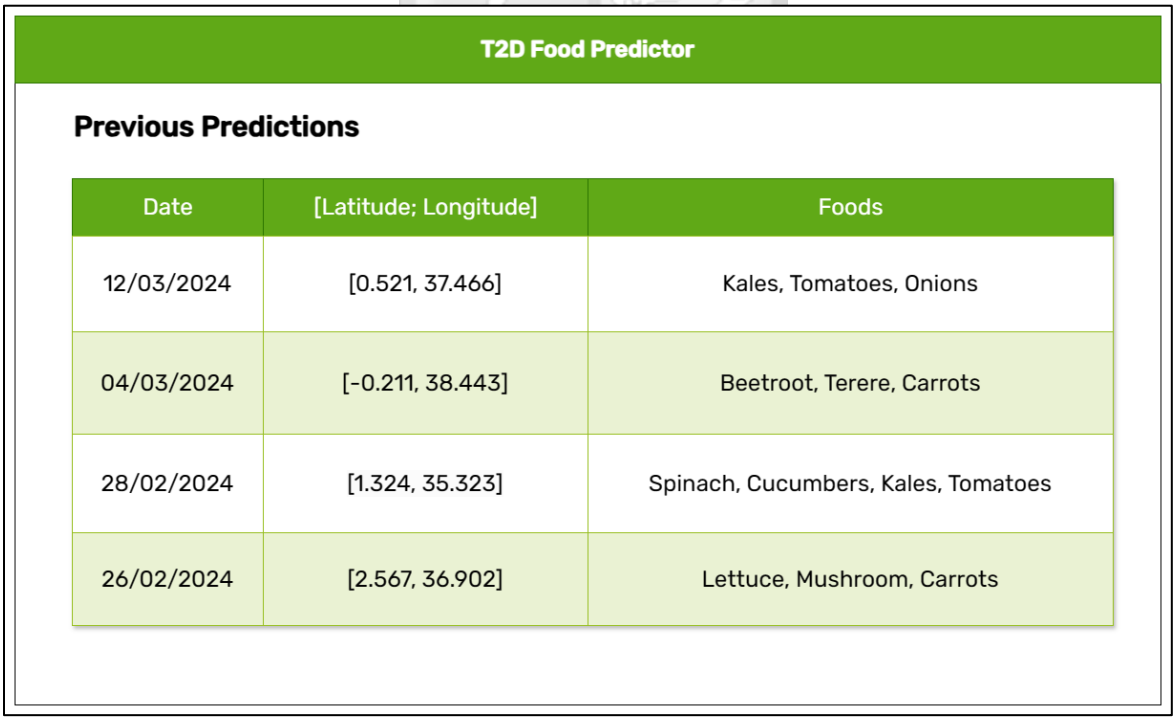


Figure 4.10: Previous Recommendations Table in Food Determinant Tool

Chapter 5: System Implementation and Testing

5.1. Introduction

This chapter provides an insight into the implementation of the developed tools for predicting soil nutrients and determining suitable food items based on soil nutrient value. It is divided into four major sections that cover development of the soil maps, classification of vegetables for recommendation, prediction of suitable vegetables based on category, and the testing of each module.

5.2. Digital Soil Nutrient Mapping

5.2.1. Development Environment

Table 5.1. describes the set of software and libraries used in developing the soil nutrient maps.

Software/Library	Version	Function
QGIS	3.34.3-Prizren	Spatial data preparation and visualisation
R	4.2.1	Spatial data manipulation and development of predictive models
RStudio	2023.09.1	R Development
GDAL	3.8.3	QGIS extension for spatial data preparation
raster	3.6-26	R library for geographic data analysis and modelling
sp	0.6-4	R library for spatial data manipulation
terra	1.7-65	R library for analysis of spatial data
corrplot	0.92	R library for visualisation of correlation matrix
psych	2.3.12	R library for correlation coefficient calculation
expss	0.11.6	R library for data manipulation
caret	6.0-94	R library for classification and regression training
ggplot2	3.4.4	R library for data visualisation
e1071	1.7-14	R library for machine learning modelling
automap	1.1-9	R library for automatic interpolation
gstat	2.1-1	R library for geostatistical modelling

Table 5.1: Development Environment for Digital Soil Nutrient Mapping

5.2.2. Data Preparation

The soil dataset applied in this study is derived from the Africa Soil Profiles Database (Leenaars et al., 2014). It includes sets of relations containing attributes on location, soil profile information and soil property values. This is illustrated in figure 5.1. that documents the database schema.

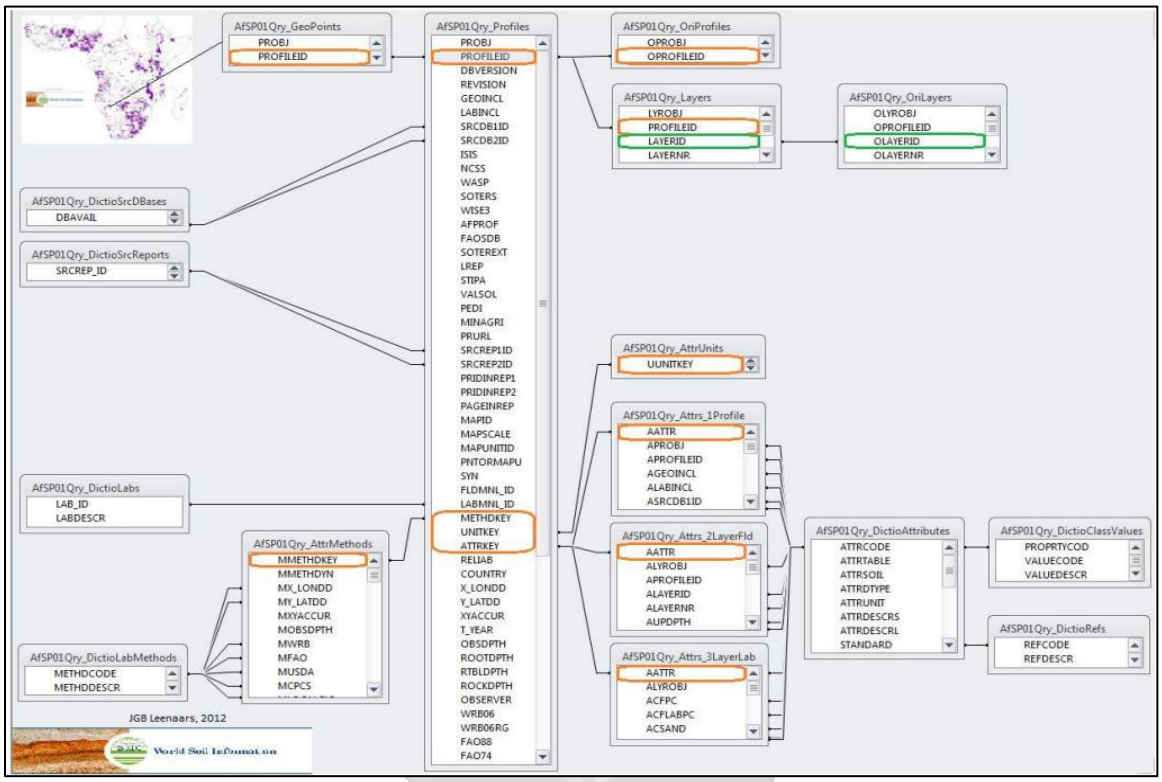


Figure 5.1: Africa Soil Profiles Database Schema (Source: Leenaars et al., 2014)

The relations of interest in this study were the 'AfSP012Qry_Layers' that contained the soil layer of interest set by depth collected and 'AfSP012Qry_Profiles' that had information on soil profiles. The following steps were carried out to derive soil profiles from Kenya:

i. Select only columns of exchangeable nutrients from AfSP012Qry_Profiles through the 'Fields' option in the QGIS Layer Properties function.

ii. Select profiles in Kenya from column Country from both relations through the expression: "Country" = 'KE' in the QGIS Select by Expression function

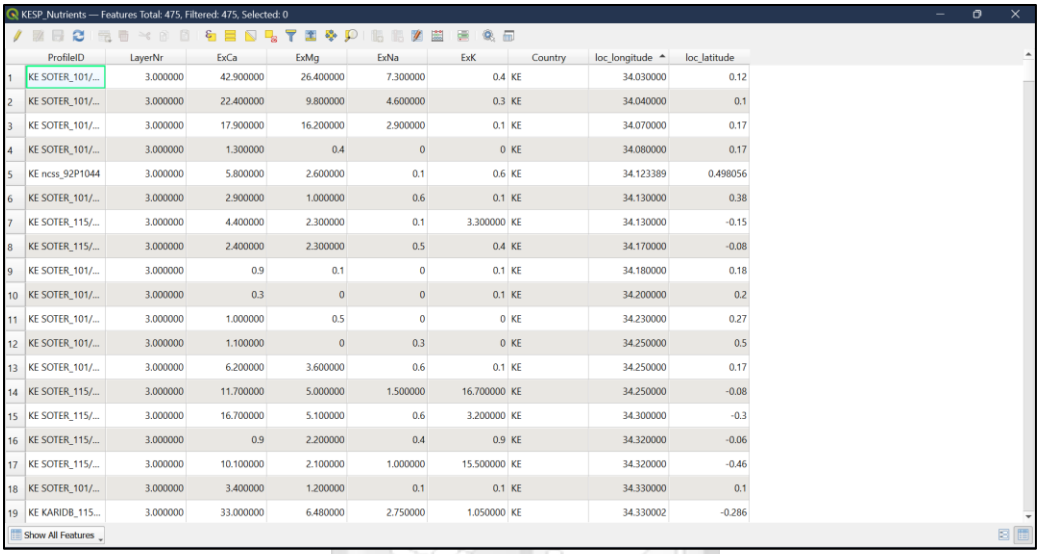
iii. Select layer number 3 from AfSP012Qry_Layers through the expression: "LayerNr" = 3 in the QGIS Select by Expression function. This layer covers up to a depth of 15cm, recommended to capture topsoil content which plants exchange nutrients from soil (Bowen et al., 2005; Hengl, Jesus, et al., 2017).

IT OMNES UNUM SINT

iv. Remove rows with NULL latitude and longitude values in 'AfSP012Qry_Profiles' through the expression: '"longitude" = 0 and "latitude" = 0' in the QGIS Select by Expression function.

v. Do Join of 'AfSP012Qry_Profiles' and 'AfSP012Qry_Layers' by Profile ID through the 'Add Vector Join' QGIS function to match location and soil nutrients.

This resulted in a table of 475 soil profiles from Kenya as depicted in figure 5.2.


KESP_Nutrients — Features Total: 475, Filtered: 475, Selected: 0

	ProfileID	LayerNr	ExCa	ExMg	ExNa	ExK	Country	loc_longitude	loc_latitude
1	KE SOTER_101/...	3.000000	42.900000	26.400000	7.300000	0.4	KE	34.030000	0.12
2	KE SOTER_101/...	3.000000	22.400000	9.800000	4.600000	0.3	KE	34.040000	0.1
3	KE SOTER_101/...	3.000000	17.900000	16.200000	2.900000	0.1	KE	34.070000	0.17
4	KE SOTER_101/...	3.000000	1.300000	0.4	0	0	KE	34.080000	0.17
5	KE ncss_92P1044	3.000000	5.800000	2.600000	0.1	0.6	KE	34.123389	0.498056
6	KE SOTER_101/...	3.000000	2.900000	1.000000	0.6	0.1	KE	34.130000	0.38
7	KE SOTER_115/...	3.000000	4.400000	2.300000	0.1	3.300000	KE	34.130000	-0.15
8	KE SOTER_115/...	3.000000	2.400000	2.300000	0.5	0.4	KE	34.170000	-0.08
9	KE SOTER_101/...	3.000000	0.9	0.1	0	0.1	KE	34.180000	0.18
10	KE SOTER_101/...	3.000000	0.3	0	0	0.1	KE	34.200000	0.2
11	KE SOTER_101/...	3.000000	1.000000	0.5	0	0	KE	34.230000	0.27
12	KE SOTER_101/...	3.000000	1.100000	0	0.3	0	KE	34.250000	0.5
13	KE SOTER_101/...	3.000000	6.200000	3.600000	0.6	0.1	KE	34.250000	0.17
14	KE SOTER_115/...	3.000000	11.700000	5.000000	1.500000	16.700000	KE	34.250000	-0.08
15	KE SOTER_115/...	3.000000	16.700000	5.100000	0.6	3.200000	KE	34.300000	-0.3
16	KE SOTER_115/...	3.000000	0.9	2.200000	0.4	0.9	KE	34.320000	-0.06
17	KE SOTER_115/...	3.000000	10.100000	2.100000	1.000000	15.500000	KE	34.320000	-0.46
18	KE SOTER_101/...	3.000000	3.400000	1.200000	0.1	0.1	KE	34.330000	0.1
19	KE KARIDB_115...	3.000000	33.000000	6.480000	2.750000	1.050000	KE	34.330002	-0.286

Show All Features

Figure 5.2: Derived Soil Profiles for Kenya

Their distribution across the country was mapped against the administrative map whose shapefiles were derived from IEBC (2018); resulting in figure 5.3.

VT OMNES VNVM SINT

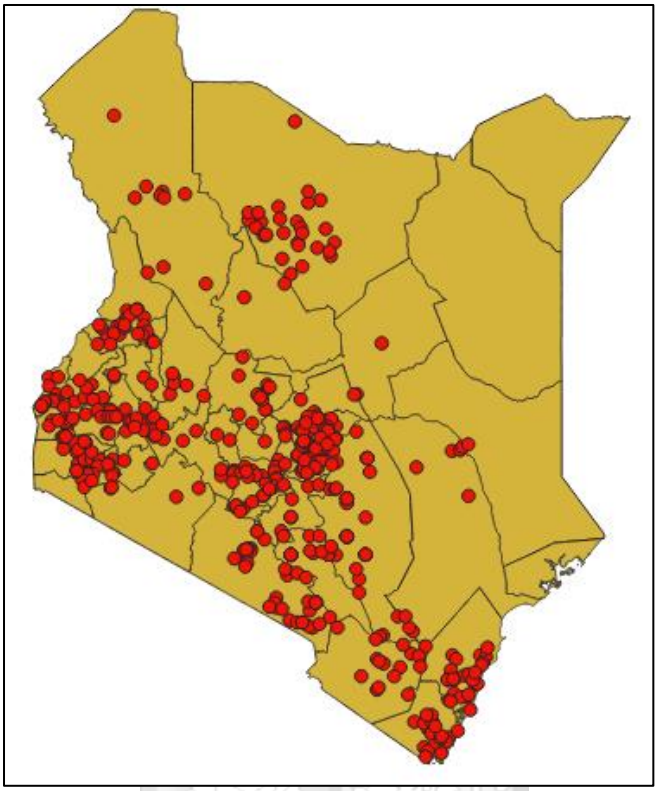


Figure 5.3: Distribution of Soil Profiles across Kenya

The researcher then identified a criterion to select a region of study: counties that have the highest representation of soil profiles and are geographically close to each other, to enable spatial continuity when modelling. This was achieved through a 'points-in-polygon' operation to derive the profile count per county, whose results for the first 15 counties are in figure 5.4. 'num_sp' represents the number of profiles per county, with the Embu and Tharaka-Nithi counties selected as they meet the criteria. Code used to run the operation is described in appendix C.1.

pts_per_cty — Features Total: 47, Filtered: 47, Selected: 2

	Shape_Leng	Shape_Area	ADM1_EN	num_sp
1	3.38665284863	0.22941354898	Embu	39.000000000000000
2	9.07416928535	2.47533754723	Kitui	36.000000000000000
3	7.41283468920	1.77918299674	Kajiado	34.000000000000000
4	11.99536214490	6.18518986668	Marsabit	28.000000000000000
5	5.74454270189	1.02512724116	Kilifi	24.000000000000000
6	5.42609734564	0.6732554386	Kwale	22.000000000000000
7	4.40059884792	0.5680862273	Meru	22.000000000000000
8	2.90378112362	0.28600479868	Siaya	18.000000000000000
9	3.17194107179	0.20968321423	Tharaka-Nithi	17.000000000000000
10	2.63239651938	0.20519376604	Murang'a	16.000000000000000
11	5.67784524415	1.39419800778	Taita Taveta	16.000000000000000
12	5.38167389936	0.77480691494	Laikipia	14.000000000000000
13	5.98377354215	0.66460275607	Makueni	13.000000000000000
14	3.15041587604	0.23038408537	Nandi	13.000000000000000
15	3.37378072663	0.3844442995	Homa Bay	12.000000000000000

Figure 5.4: Points-in-Polygon Results

The extent for this region was selected and exported as a shapefile through the QGIS "Export" function. Soil profiles from this region were then extracted and stored as spatial points, and attributes on exchangeable potassium and exchangeable magnesium retained through the "Extract by location" operation. This process is demonstrated in appendix C.2. The second set of input on environmental covariates was selected from SoilGrids250m (Hengl, Jesus, et al., 2017), that capture different covariate attributes at a resolution of 250m. Table 5.2 describes the covariates selected for modelling of exchangeable potassium and exchangeable magnesium.

Property	Code	Name
Climate	B07CHE3	Temperature Annual Range
Organism	BARL10	Bare ground
Relief	DEMENV5	Land surface elevation
Organism	MAXENV3	EVI - enhanced vegetation index
Climate	PRSCHE3	Total annual precipitation
Relief	SLPMRG5	Terrain slope
Relief	TPIMRG5	Topographic position index

Organism	TREL10	Tree cover
Climate	TWIMRG5	SAGA Wetness Index
Relief	VDPMRG5	Valley depth
Climate	B13CHE3	Precipitation of wettest month
Climate	B14CHE3	Precipitation of driest month
Relief	EXTGSW7	Surface water extent
Climate	B04CHE3	Temperature Seasonality

Table 5.2: Environmental Covariates selected for the Soil Mapping

Each covariate was clipped to fit the region of study using the 'Clip raster by extent' function, which is demonstrated in appendix C.3. More data points were added to capture the region that would be encapsulated by the covariates after clipping. This resulted in a model set of 76 data points distributed within the rectangle as in figure 5.5. The polygons in pink represent the counties selected for mapping.

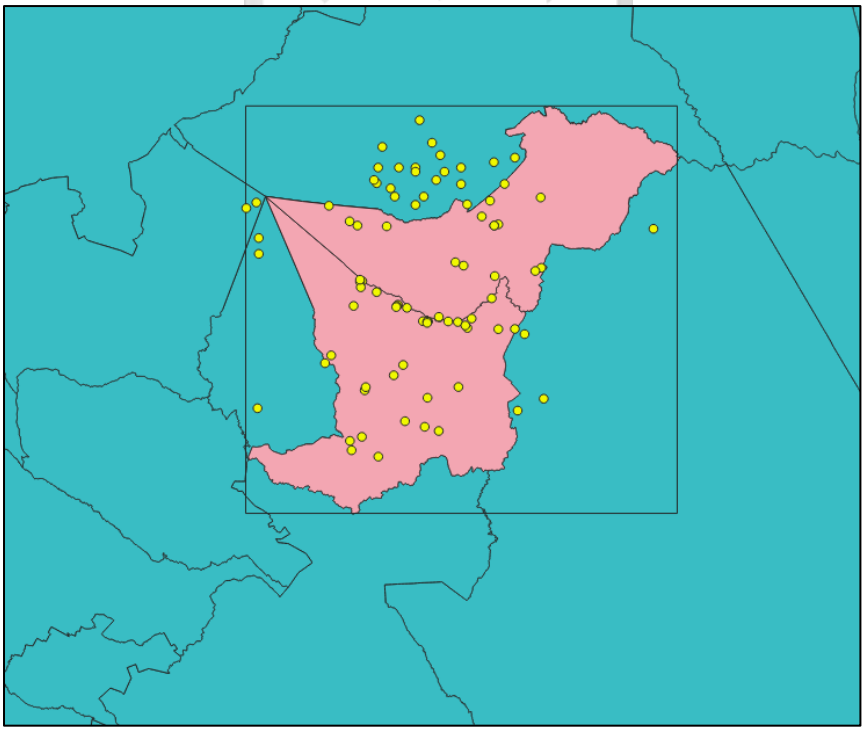


Figure 5.5: Selected Soil Profiles for Modelling

5.2.3. Exploratory Data Analysis

Figure 5.6 demonstrates summary statistics of soil profiles against the covariates, providing an insight into the data cleaning strategies. This was generated through the `summary()` function in R.

```
##        X               Y            ProfileID             ExMg
##  Min.   :37.26   Min.   :-0.7778   Length:87          Min.   :-9999.00
##  1st Qu.:37.57   1st Qu.:-0.4641   Class :character   1st Qu.:    0.35
##  Median :37.69   Median :-0.3542   Mode  :character   Median :    1.30
##  Mean   :37.68   Mean   :-0.3451                      Mean   :-1376.61
##  3rd Qu.:37.79   3rd Qu.:-0.1622                      3rd Qu.:    3.00
##  Max.   :38.24   Max.   : 0.0344                      Max.   :   18.60
##       ExK             annual_ppt       bareground       driest_ppt
##  Min.   :-9999.00   Min.   : 600.8   Min.   : 0.000   Min.   : 3.373
##  1st Qu.:    0.09   1st Qu.:1131.9   1st Qu.: 0.000   1st Qu.:17.121
##  Median :    0.20   Median :1394.3   Median : 3.000   Median :22.768
##  Mean   :-1378.79   Mean   :1382.0   Mean   : 5.161   Mean   :24.567
##  3rd Qu.:    0.65   3rd Qu.:1618.6   3rd Qu.: 7.500   3rd Qu.:32.367
##  Max.   :    2.00   Max.   :2157.5   Max.   :33.000   Max.   :61.313
##  land_elevation sfc_water_extent temp_annual_range temp_seasonality
##  Min.   : 516   Min.   :0        Min.   :14.78     Min.   :60.85
##  1st Qu.: 865   1st Qu.:0        1st Qu.:15.66     1st Qu.:78.60
##  Median :1151   Median :0        Median :15.93     Median :81.06
##  Mean   :1330   Mean   :0        Mean   :15.96     Mean   :81.23
##  3rd Qu.:1518   3rd Qu.:0        3rd Qu.:16.28     3rd Qu.:84.32
##  Max.   :4379   Max.   :0        Max.   :16.97     Max.   :90.98
##  terrain_slope    topo_pst_index        treecover      valley_depth
##  Min.   : 1.000   Min.   :-283.0000   Min.   : 0.00   Min.   :-1769.0
##  1st Qu.: 2.000   1st Qu.: -57.0000   1st Qu.: 8.00   1st Qu.:  192.5
##  Median : 4.000   Median :  -5.0000   Median :15.00   Median : 1990.0
##  Mean   : 6.563   Mean   :   0.6207   Mean   :20.71   Mean   : 2016.3
##  3rd Qu.: 9.000   3rd Qu.:  70.5000   3rd Qu.:25.00   3rd Qu.: 3429.0
##  Max.   :22.000   Max.   : 293.0000   Max.   :92.00   Max.   : 7297.0
##    veg_index    wetness_index     wettest_ppt
##  Min.   : 837   Min.   : 68.00   Min.   :135.7
##  1st Qu.:1228   1st Qu.: 91.50   1st Qu.:256.4
##  Median :1477   Median : 99.00   Median :305.7
##  Mean   :1514   Mean   : 98.03   Mean   :305.2
##  3rd Qu.:1745   3rd Qu.:105.00   3rd Qu.:357.0
##  Max.   :2475   Max.   :120.00   Max.   :455.5
```

Figure 5.6: Summary Statistics of Soil Profiles and Environmental Covariates

The following processes were carried out to clean the data:

i. Removal of -9999.99 and 0.00 values from nutrient columns by use of These were imputed by the mean for the respective column, which is statistically sound in spatial data analysis (Baker et al., 2014). This is presented in equation 5.1 for exchangeable Magnesium.

$$sp\_covs\$ExMg <- na\_if(sp\_covs\$ExMg, -9999.00)$$
$$sp\_covs\$ExMg[is.na(sp\_covs\$ExMg)] <- mean(sp\_covs\$ExMg, na.rm = T)$$

Equation 5.1: R Code for Detecting and Imputing Null Values

ii. Dropping null columns and rows, resulting in the surface water extent being dropped. This implied that the region under study is largely ground without much significant water bodies. This is as presented in equation 5.2.

$$sp\_covs <- subset(sp\_covs, select = -c(sfc\_water\_extent))$$
$$sp\_covs <- sp\_covs[complete.cases(sp\_covs),]$$

Equation 5.2 R Code for Dropping Null Columns and Rows

Histograms of nutrients under study were then generated through the `hist()` function in R. The major inference made was magnesium having relatively larger amounts across the region compared to potassium even in its exchangeable format. These histograms are demonstrated in figures 5.7 and 5.8.

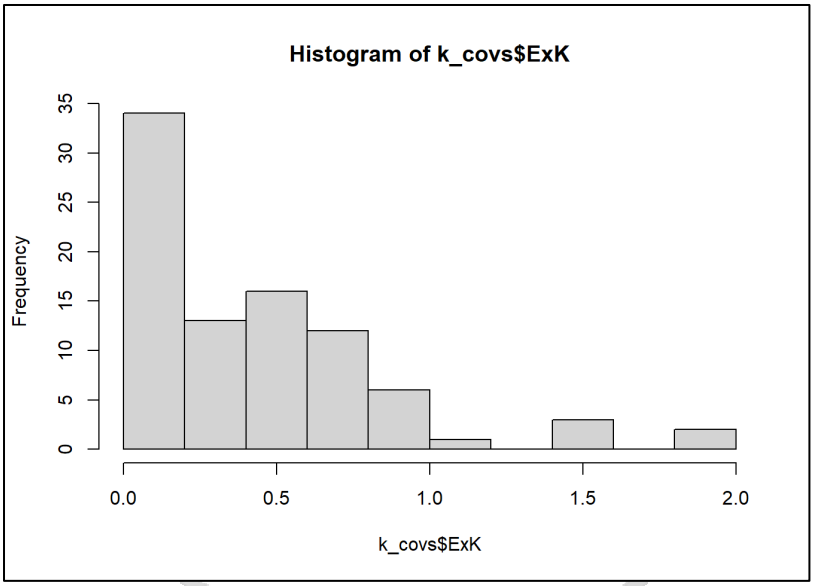


Figure 5.7: Histogram for Distribution of Exchangeable Potassium

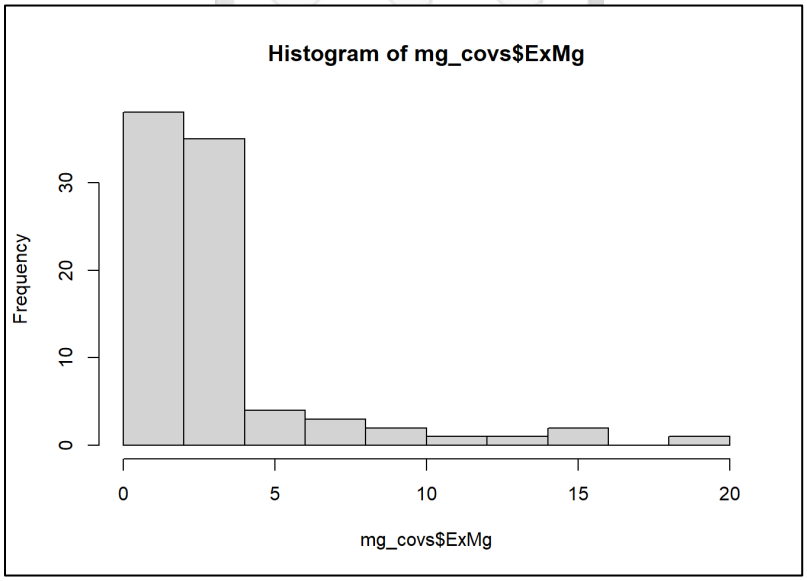


Figure 5.8: Histogram for Distribution of Exchangeable Magnesium

Correlation tests between individual nutrients and covariates were then performed using Pearson's correlation coefficient, due to small variations between them. This was done by the code in equation 5.3 that led to the figures 5.9 and 5.10 as the resulting matrices. Figure 5.11 illustrates actual scores in absolute values achieved from the correlation tests, generated by the code in equation 5.4.

$$corrplot(cor(k\_covs[4{:}17], method = "pearson"), method = "color")$$
$$corrplot(cor(mg\_covs[4{:}17], method = "pearson"), method = "color")$$

Equation 5.3: Correlation Plots Generation

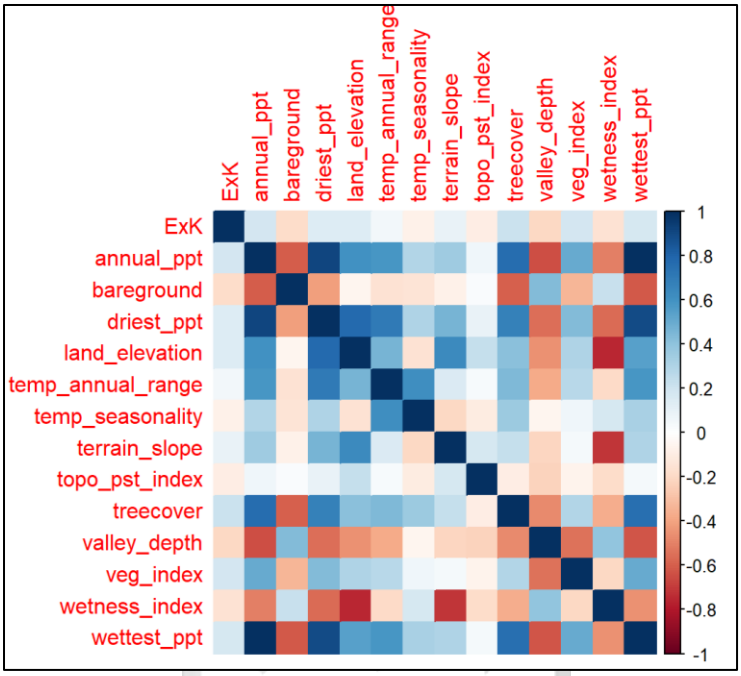


Figure 5.9: Correlation Matrix for Exchangeable Potassium

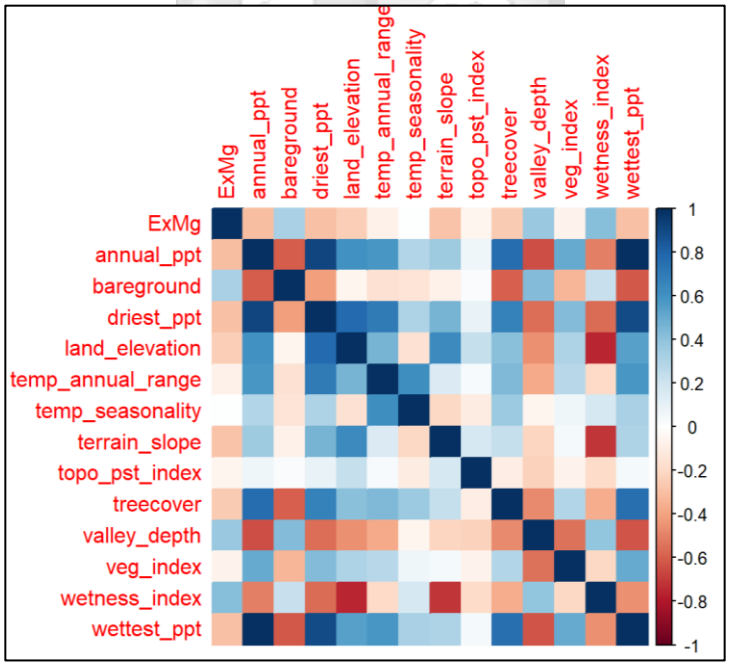


Figure 5.10: Correlation Matrix for Exchangeable Magnesium

$$k\_cor <- abs(round(cor(x = as.matrix(k\_covs[,4]), y = as.matrix(k\_covs[,c(5:17)], method = "pearson")),2))$$

$$mg\_cor <- abs(round(cor(x = as.matrix(mg\_covs[,4]), y = as.matrix(mg\_covs[,c(5:17)], method = "pearson")),2))$$

Equation 5.4: R Code to Retrieve Correlation Indices

```
k_cor

##        annual_ppt bareground driest_ppt land_elevation temp_annual_range
## k_corr       0.18       0.19       0.15           0.15              0.05
##        temp_seasonality terrain_slope topo_pst_index treecover valley_depth
## k_corr             0.08          0.09           0.09      0.21          0.2
##        veg_index wetness_index wettest_ppt
## k_corr      0.19          0.16        0.17

mg_cor

##         annual_ppt bareground driest_ppt land_elevation temp_annual_range
## mg_corr       0.31       0.32        0.3           0.25              0.07
##         temp_seasonality terrain_slope topo_pst_index treecover valley_depth
## mg_corr             0.01          0.29           0.06      0.25         0.37
##         veg_index wetness_index wettest_ppt
## mg_corr      0.07          0.42        0.29
```

Figure 5.11: Correlation Test Results

All covariates presented relatively weak correlations to both potassium and magnesium, therefore covariates that presented scores of greater than 0.15 with potassium and 0.2 with magnesium were selected, leading to these covariates being applied in modelling:

i. Exchangeable Potassium: Annual precipitation, Bare ground, Precipitation of driest month, Land surface elevation, enhanced vegetation index, Tree cover, SAGA wetness index, Precipitation of wettest month

ii. Exchangeable Magnesium: Annual precipitation, Bare ground, Precipitation of driest month, Land surface elevation, terrain slope, Valley depth, Tree cover, SAGA wetness index, Precipitation of wettest month

5.2.4. Modelling Process and Results

Since both nutrient values were right skewed in their distribution, as per figures 5.7 and 5.8, they were log transformed to create a more uniform distribution for modelling. This resulted in the distributions depicted in figures 5.12 and 5.13. Scripts used to perform histogram generation and log transformation are described in equation 5.5.

$$hist(k\_dat\$ExK)$$
$$k\_dat\$log\_ExK \ <- log(k\_dat\$ExK)$$

Equation 5.5: Histogram Generation and Log Transformation of Nutrient Distributions

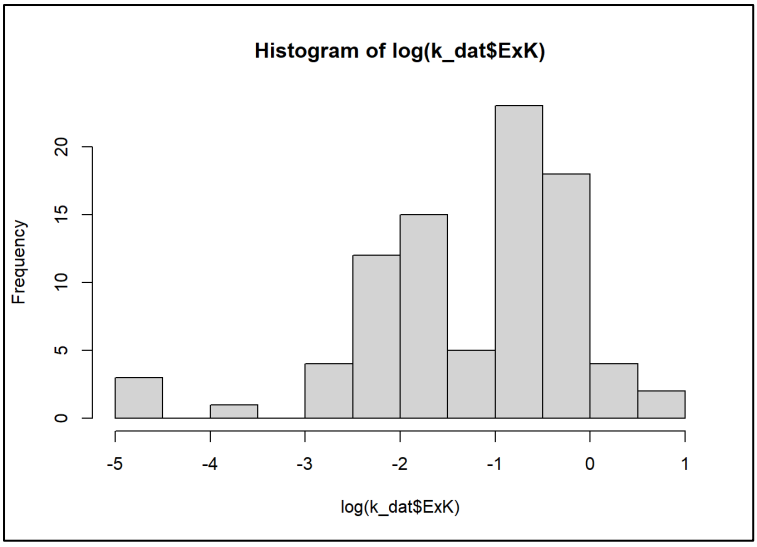


Figure 5.12: Histogram of Log Transformed Exchangeable Potassium

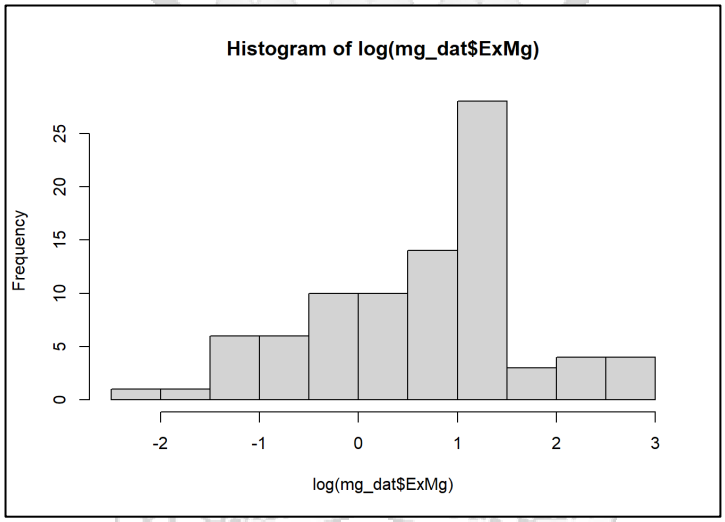


Figure 5.13: Histogram of Log Transformed Exchangeable Magnesium

Data was then randomly split into an 80:20 ratio for training and testing the models. Code for this function is provided in appendix C.4. Equations for regression were created for the respective nutrients for fitting in the MLR, SVM, and MLRK models. The equations in R code for generating the magnesium maps are as in equations 5.6 and 5.7, while those for modelling by MLR and SVM are illustrated in equations 5.8. and 5.9

$$mg\_regeqn <- as.formula(log\_ExMg \sim annual\_ppt + bareground + driest\_ppt + land\_elevation + terrain\_slope + treecover + valley\_depth + wetness\_index + wettest\_ppt)$$

Equation 5.6: R Regression Equation for Exchangeable Magnesium Map

$$k\_regeqn <- as.formula(log\_ExK \sim annual\_ppt + bareground + land\_elevation + treecover + valley\_depth + veg\_index + wetness\_index + wettest\_ppt)$$

Equation 5.7: R Regression Equation for Exchangeable Potassium Map

$$k\_mlr <- train(k\_regeqn, data = k\_train, method = "lm")$$

$$mg\_mlr <- train(mg\_regeqn, data = mg\_train, method = "lm")$$

Equation 5.8: R Code for Multiple Linear Regression

$$k\_svm <- train(k\_regeqn, data = k\_train, method = 'svmRadial')$$
$$mg\_svm <- train(mg\_regeqn, data = mg\_train, method = 'svmRadial')$$

Equation 5.9: R Code for Support Vector Machines

In SVM, the hyperparameters were fine-tuned to increase modelling accuracy, as prescribed in section 2.3.3 and in (Suleymanov et al., 2021). This involved tuning the gamma and cost hyperparameters and determining the optimal values as captured in equation 5.9 which captures tuning and parameter derivation for SVM modelling of the exchangeable potassium map.

$$svm\_tune <- tune(svm, k\_regeqn, data = k\_train[, c("log\_ExK", names(k\_covs))], ranges = list(gamma = seq(0.2,0.5,0.05), cost = seq(0.1,1,0.1)))$$

$$svm\_tune\$best.parameters$$

Equation 5.10: R Code for SVM Hyperparameter Tuning

This was then fit to the SVM model as in equation 5.11.

$$hp <- expand.grid(C = svm\_tune\$best.parameters\$cost, sigma = svm\_tune\$best.parameters\$gamma)$$

$$k\_svm <- train(k\_regeqn, data = k\_train, method = 'svmRadial', tuneGrid = hp)$$

Equation 5.11: R Code for Optimum SVM Tuning

Modelling by ordinary kriging was done by equations 5.12 and 5.13 then applied to the automatic interpolation function `autoKrige()` as in equation 5.14. Before kriging, the data was converted into a spatial format to allow for estimation of spatial patterns. The technique used in conversion is available in appendix C.5.

$$mg\_okeqn <- as.formula(log\_ExMg \sim 1)$$

Equation 5.12: R Ordinary Kriging Equation for Exchangeable Magnesium Map

$$k\_okeqn <- as.formula(log\_ExK \sim 1)$$

Equation 5.13: R Ordinary Kriging Equation for Exchangeable Potassium Map

$$mg\_ok <- autoKrige(mg\_okeqn, input\_data = mg\_train, verbose = TRUE)$$
$$k\_ok <- autoKrige(k\_okeqn, input\_data = k\_train, verbose = TRUE)$$

Equation 5.14: Modelling by Ordinary Kriging

Modelling by regression kriging was also performed using the `autoKrige()` function, where the regression equation was applied, as in equation 5.15.

$$mg\_mlrk <- autoKrige(mg\_regeqn, new\_data = mg\_covs, input\_data = mg\_train, verbose = TRUE)$$
$$k\_mlrk <- autoKrige(k\_regeqn, new\_data = mg\_covs, input\_data = k\_train, verbose = TRUE)$$

Equation 5.15: Modelling by Regression Kriging

The generated maps from these processes is illustrated in figures 5.14 and 5.15, set to the extent of the region of study.

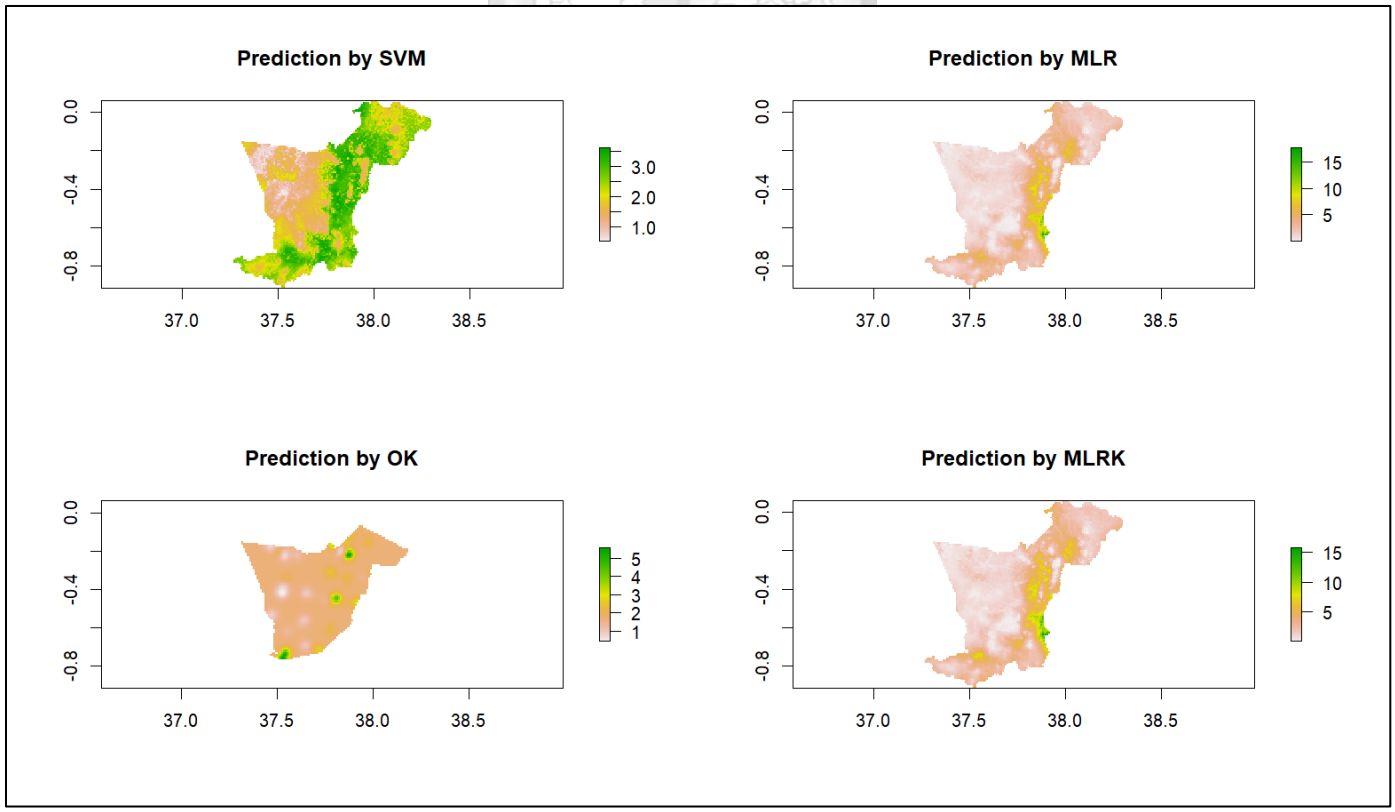


Figure 5.14: Predicted Exchangeable Magnesium Maps

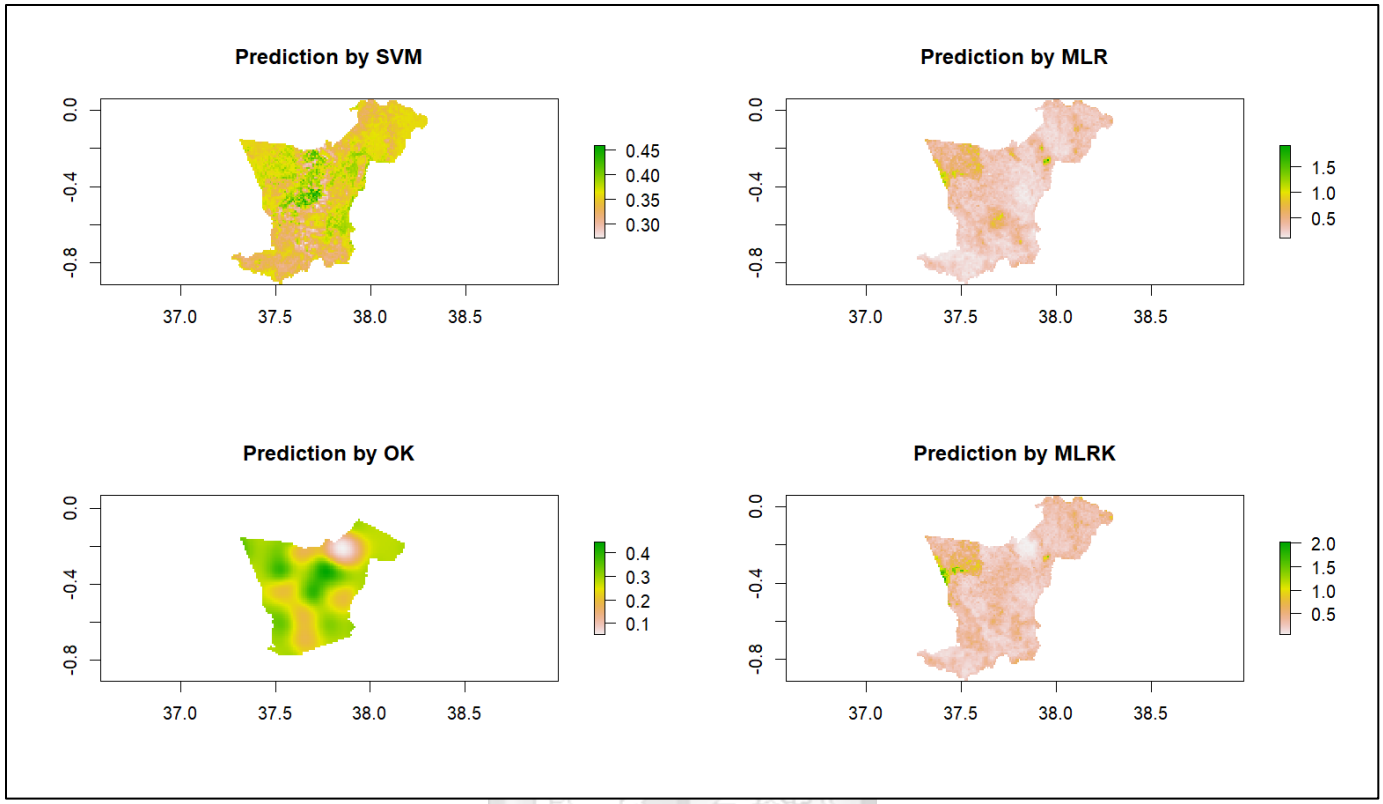


Figure 5.15: Predicted Exchangeable Potassium Maps

5.3. Vegetable Classification

5.3.1. Development Environment

Table 5.3. illustrates the software and libraries used to carry out classification of vegetable foods in preparation for prediction by the predictive tool.

Software/Library	Version	Function
Python	3.10.4	Data preparation, modelling, and visualisation
R	4.2.1	Exploratory data analysis, data preparation
RStudio	2023.09.1	R Development
Visual Studio Code	1.87.2	Python development
fs	1.5.2	R library for file operations
readxl	1.4.1	R library for manipulating Excel files
dplyr	1.0.10	R library for data manipulation
pandas	2.1.4	Python library for data manipulation
seaborn	0.13.1	Python library for data visualisation

matplotlib	3.8.2	Python library for data visualisation
scikit-learn	1.3.2	Python library for machine learning tools (Pedregosa et al., 2011)
pickle	3.8	Python library for object serialisation (model import and export)

Table 5.3: Development Environment for Vegetable Classification

5.3.2. Dataset Description

The food dataset applied in this study is the Kenya Food Composition Tables (FAO & Government of Kenya, 2018) that consists of one major relation of interest the 'UserDatabase-KFCT2018' that contains a list of foods consumed in the country and their respective nutritional content. A snapshot of this is depicted in figure 5.16.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Food Identification			Energy, proximates, minerals and vitamins								Minerals				
2	(All values expressed per 100 g of Edible Portion on Fresh Weight Basis (EP)															
3	Code	Food name	Edible conversion factor	Energy (kJ)	Energy (kcal)	Water (g)	Protein (g)	Fat (g)	Carbohydrate available (g)	Fibre (g)	Ash (g)	Ca (mg)	Fe (mg)	Mg (mg)	P (mg)	K (mg)
4																
5		1 CEREALS AND CEREAL PRODUCTS														
569	5033	Raspberry, raw	1	226	54	85.2	1.4	0.2	8.9	5.0	0.7	27	0.8	25	37	150
570	SD or min-max					1.9	1-1.8				0.4-0.9	14-40	0.5-1.1		34-39	
571	n					3	2	1		1	2	2	2	1	2	1
572	5034	Strawberry, raw	0.96	146	35	90	0.7	0.2	6.3	2.5	0.4	19	0.6	8	27	158
573	SD or min-max					87.8-92.1						8-30			24-30	
574	n					2	1	1		1	1	2	1	1	2	1
575	5035	Tangerine, juice, non-commercial	1	195	46	88.3	0.6	0.2	10.3	0.2	0.4	12	0.2	11	20	178
576	SD or min-max					86.1-90.5	0.2				0.1	8-18				
577	n					2	3	1		1	3	2	1	1	1	1
578	5036	Tangerine, pulp, raw	0.75	218	52	86.2	0.87	0.2	10.8	1.5	0.4	26	0	17	17	150
579	SD or min-max						0.7-1				0.3-0.5					
580	n					1	2	1		1	2	1	1	1	1	1
581	5037	Tree tomato, dark red skin, peeled, raw	0.4	177	43	85.6	2.1	[0.3]	4.6	6.5	0.9	56	1	19	51	133
582	SD or min-max															
583	n					1	1	1		1	1	1	1	1	1	1

04 USERDATABASE-KFCT2018 | 05 Yield-factors | 06 Retention Factors | 07 Index of f

Figure 5.16: Kenya Food Composition Tables Dataset

To retrieve data specific to this study, the following data cleaning processes were conducted:

i. Slicing out rows of foods under the 'Vegetables' category and selecting the 'Code'. 'Food name', 'Magnesium', and 'Potassium' columns.
ii. Removing rows with null data as well as duplicate items by name of food.
iii. Removing cooked vegetables. This was through the str_detect() function in R that checked for the words 'canned', 'boiled', 'steamed', 'stewed', 'grilled' or 'baked' in the food name.
iv. Converting nutrient values to float datatype since they were saved as strings.

This resulted in a dataset of 36 points that feature nutrient values of raw vegetables. The data cleaning was done through the code in equation 5.16.

$$raw\_foods <- raw\_foods[!(is.na(raw\_foods\$Vegetable) | raw\_foods\$Vegetable == ""),]$$

$$raw\_foods <- raw\_foods[!duplicated(raw\_foods\$Vegetable),]$$

$$raw\_foods$$

$$<- raw\_foods[!str\_detect(raw\_foods\$Vegetable, 'canned|boiled|steamed|stewed|grilled|baked'),]$$

$$raw\_foods <- raw\_foods \%>\% mutate\_at(c('Calcium','Magnesium','Potassium','Sodium'), as.numeric)$$

Equation 5.16: R Code for Vegetable Data Processing

5.3.3. Exploratory Data Analysis

In preparation for clustering, the distribution of vegetables against magnesium and potassium values was reviewed. This was achieved by running the Python script as in equation 5.17. Figure 5.17 indicates that most vegetables overall have low magnesium but high potassium content. Most vegetables have characteristically low potassium and magnesium content.

$$sns.scatterplot(data = foods, x = 'Magnesium', y = 'Potassium')$$

Equation 5.17: Python Code for Visualising Distribution of Vegetables

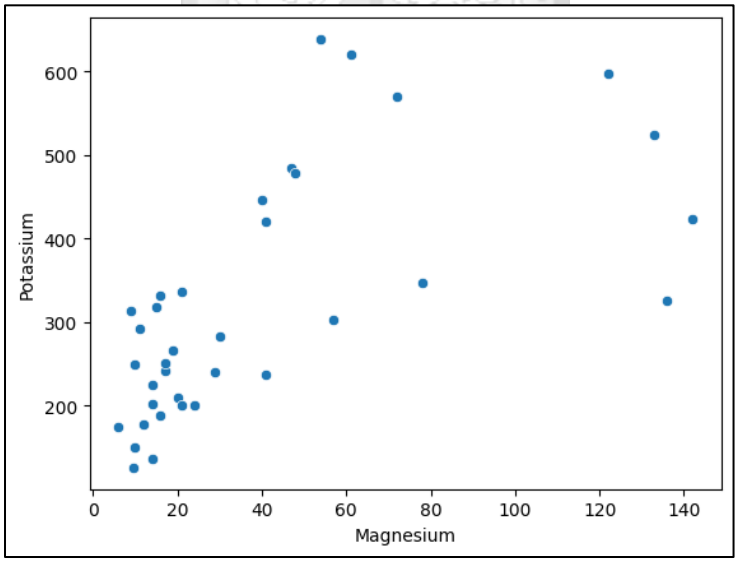


Figure 5.17: Distribution of Vegetables against Potassium and Magnesium Content

5.3.4. Modelling Process and Results

The algorithm selected for categorisation was k-means clustering, due to its simplicity and flexibility (Ikotun et al., 2023). In preparation for this, the nutrient values were normalised by the StandardScaler() function in scikit-learn, as depicted in figure 5.18. This is to avoid effects of outliers in modelling.

```
# normalise nutrient values
scaler = StandardScaler()
nutrients = scaler.fit_transform(nutrients)
nutrients
```

```
array([[ 2.17009825,  1.95691931],
       [-0.51700368, -0.78095977],
       [ 0.03622319,  0.71178567],
       [-0.49065954,  0.11044272],
       [-0.78044505, -0.49797485],
       [-0.27990645, -0.56164645],
       [-0.80678918, -0.0522736 ],
       [-0.78044505, -1.19836252],
       [-0.88582159, -1.03564619],
       [-0.59603609, -0.55457183],
       [-0.59603609, -0.49090023],
       [-0.75410091, -0.20084068],
       [-0.54334782, -0.37770626],
       [-0.62238023, -0.93660147],
       [ 0.19428801,  1.16456154],
       [-0.6750685 , -1.30448186],
```

Figure 5.18: Nutrient Normalisation for Clustering

Clustering was done with the elbow metric under consideration, to determine the most suitable number of clusters. When plotted, the point at which the elbow plot registers a sharp decline is considered the optimal number of clusters (Syakur et al., 2018). In this study, the elbow method registered a score of 3 as the ideal number of clusters, as depicted in figure 5.19. Code for executing the elbow plot is demonstrated in appendix C.6.

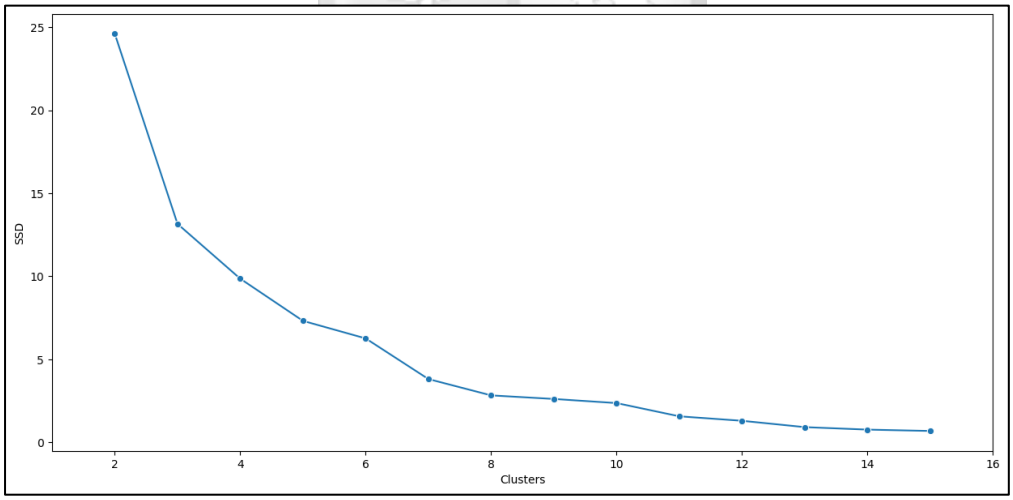


Figure 5.19: Elbow Score for K-means Clustering.

Clustering was done with the k value set to 3 and the corresponding clusters set to each food item, as demonstrated in equation 5.18. This created three categories of vegetables whose distribution is illustrated in figure 5.20.

$$km = KMeans(n\_clusters = 3, random\_state = 0, n\_init = "auto")$$
$$km.fit(nutrients)$$
$$foods['Category'] = km.labels\_$$

Equation 5.18: K-means Clustering with 3 clusters

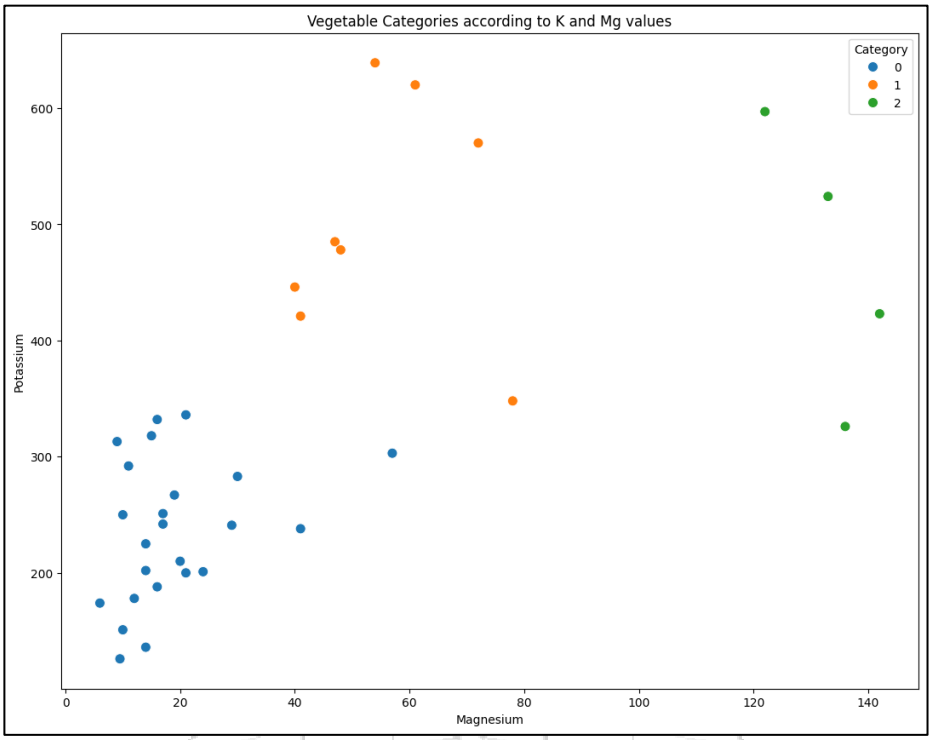


Figure 5.20: Categories of Vegetables after k-means Clustering.

5.4. Food Predictor Web Application

5.4.1. Development Environment

Table 5.3 describes the software and libraries used to develop the web application and tool for food prediction.

Software/Library	Version	Function
Python	3.10.4	Data preparation, modelling, and visualisation
Visual Studio Code	1.87.2	Python development
PostgresSQL	16.2	Data storage
pgAdmin 4	8.2	PostgresSQL server and database management

Flask	1.5.2	Python framework for building web applications
Jinja2	1.4.1	Python library to handle HTML templates for web applications
psycopg2-binary	1.0.10	Python library for handling PostgresSQL requests
os	3.10.4	Python library for operating system processes
rasterio	1.3.9	Python library for manipulating raster files
pandas	2.1.4	Python library for data manipulation
seaborn	0.13.1	Python library for data visualisation
matplotlib	3.8.2	Python library for data visualisation
scikit-learn	1.3.2	Python library for machine learning tools
pickle	3.8	Python library for object serialisation (model import and export)
Leaflet	1.9.4	JavaScript library for map interaction

Table 5.4: Development Environment for Food Predictor Web Application

5.4.2. Tool Description

The tool developed to facilitate prediction of vegetables to diabetic patients is a web application with a map to allow nutritionists to select the location, a form to generate food items based on soil nutrient value, a table of determined food items and another listing all available vegetables.

Upon accessing the interface, they are presented with a map tool to select a location and input their coordinates to access soil nutrient values. This is presented in figure 5.21, with the code for generating the map is provided in appendix C.7.

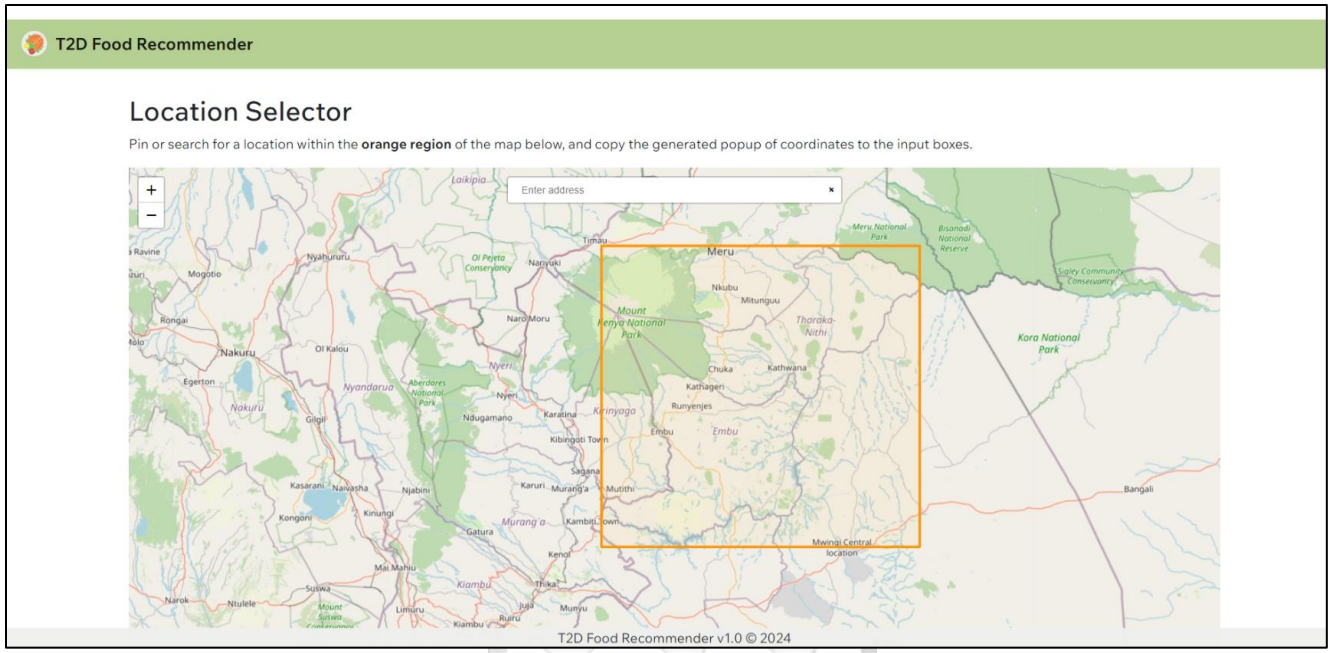


Figure 5.21: Food Predictor Tool Map Interface

Since the maps generated so far only cover Embu and Tharaka-Nithi counties, the yellow polygon is to bind the user from selecting a location outside the area bounded by soil maps. The user then inputs the coordinates into the 'Food Generator' form, select the soil nutrient of interest and generate food items, as illustrated in figure 5.22.

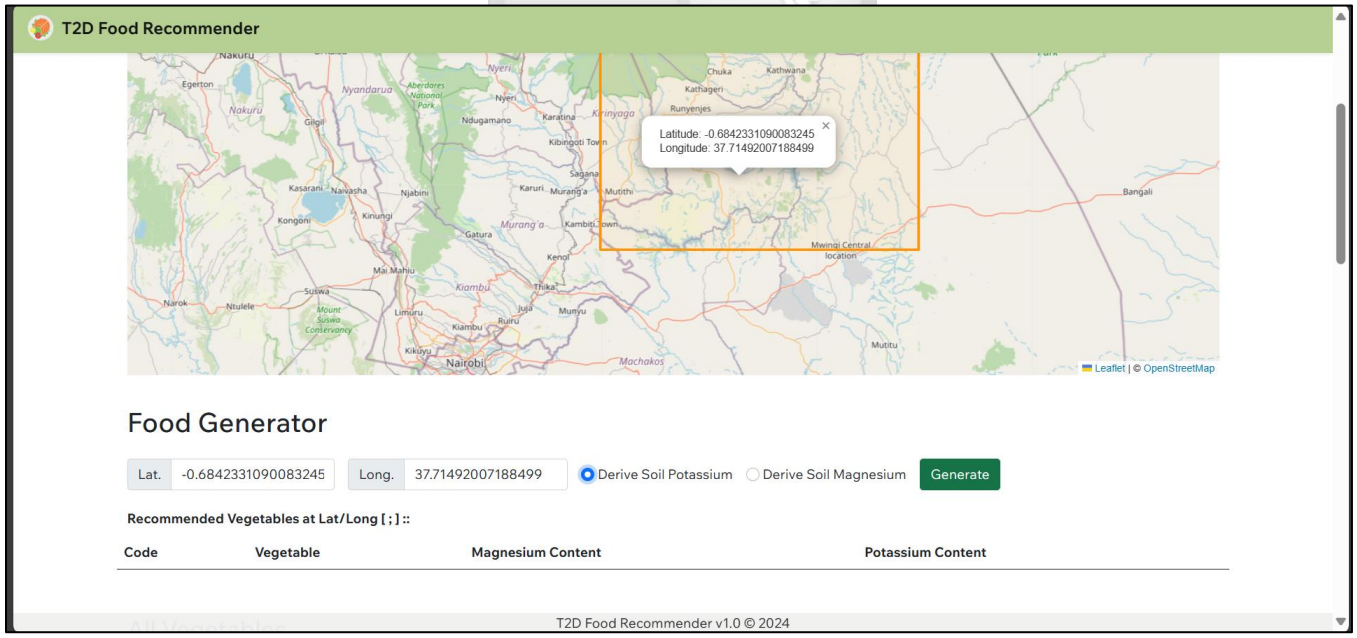


Figure 5.22: Food Generator Form for the Predictor Tool

The tool then provides a list of predicted foods and their nutrient values for the nutritionist to determine the most suitable to the patient, as illustrated in figure 5.23.

Recommended Vegetables at Lat/Long [-0.6842331090083245 ; 37.71492007188499] :: Soil Potassium: 128.86

Code	Vegetable	Magnesium Content	Potassium Content
4036	Tomato, red, ripe, raw	9.5	126.0
4016	Cucumber, green, unpeeled, raw	14.0	136.0
4008	Capsicum (sweet peper), green, raw	10.0	151.0
4009	Capsicum (sweet peper), red, raw	6.0	174.0
4027	Pumpkin, flesh, yellow w/o seeds, raw	12.0	178.0
4014	Courgette, green, unpeeled, raw	16.0	188.0
4017	Eggplant / Brinjal, different varieties, whole edible, raw	21.0	200.0
4022	Lettuce, not further defined, raw	24.0	201.0
4021	Leeks, bulb and stem, raw	14.0	202.0
4002	Bitter gourd, whole, different varieties, raw	20.0	210.0

Figure 5.23: Results of Food Prediction by the Predictor Tool

Since the soil maps are represented in exchangeable amounts i.e. exchangeable potassium and magnesium, the values retrieved from the maps are calculated by an identified conversion factor to provide an estimate of available potassium and magnesium in the soil. The researcher incorporated a conversion factor of 122 for magnesium and 390 for potassium as stated in (State of New South Wales through Local Land Services, 2020). The workflow described for food prediction is covered by the code snippet in figure 5.24.

VT OMNES VNVM SINT

```
def predict_veg(latitude, longitude, map):
    av_nutrients = []
    nut_output = dict();
    conv_mg = 122
    conv_k = 390

    latitude = float(latitude)
    longitude = float(longitude)

    # open maps
    k_map = rasterio.open(os.path.join(dir, "recommender", "static", "images", "k_map.tif"))
    mg_map = rasterio.open(os.path.join(dir, "recommender", "static", "images", "mg_map.tif"))

    # read values and convert to available nutrients
    mg_row, mg_col = mg_map.index(longitude,latitude)
    k_row, k_col = k_map.index(longitude,latitude)
    soil_mg = (mg_map.read(1)[mg_row, mg_col] * conv_mg)
    soil_k = (k_map.read(1)[k_row, k_col] * conv_k)

    av_nutrients.append({
        'Magnesium': round(soil_mg,2),
        'Potassium': round(soil_k,2)
    })
    soil_nutrients = pd.DataFrame(av_nutrients)

    # load models
    with open((os.path.join(dir, "recommender", "src", "k_dt.pkl")), 'rb') as kf:
        k_mod = pkl.load(kf)
    with open((os.path.join(dir, "recommender", "src", "mg_dt.pkl")), 'rb') as mgf:
        mg_mod = pkl.load(mgf)

    # categorise based on requested content
    if map == "potassium":
        nut_output['soil'] = "Soil Potassium: "+str(round(soil_k,2))
        category = k_mod.predict(pd.DataFrame(soil_nutrients.iloc[:,1]))
    if map == "magnesium":
        nut_output['soil'] = "Soil Magnesium: "+str(round(soil_mg,2))
        category = mg_mod.predict(pd.DataFrame(soil_nutrients.iloc[:,0]))
    nut_output['category'] = category

    return nut_output

def get_foods(category, map):
    category = category.tolist()
    cat = category.pop(0)
    db = get_db()
    cur = db.cursor(cursor_factory=RealDictCursor)
    if map == "potassium":
        cur.execute("SELECT * FROM foods.veg WHERE category=%s ORDER BY potassium ASC LIMIT 10",(cat,))
    if map == "magnesium":
        cur.execute("SELECT * FROM foods.veg WHERE category=%s ORDER BY magnesium ASC LIMIT 10",(cat,))
    vegs = cur.fetchall()
    cur.close()
    return vegs
```

Figure 5.24: Food Prediction Workflow

5.5. System Testing and Validation

5.5.1. Digital Soil Nutrient Maps

In testing accuracy of the predicted maps, the RMSE value was calculated by comparing the predicted values to the test set generated in section 5.2.4. Table 5.5 illustrates the RMSE values for both

magnesium and potassium maps. Maps generated by SVM were produced with the least amount of RMSE rendering them the most accurate.

Technique	Exchangeable Potassium	Exchangeable Magnesium
SVM	0.37	2.74
MLR	0.43	3.67
OK	0.39	2.75
MLRK	0.45	3.67

Table 5.5: Results from Map Testing

5.5.2. Vegetable Classification

To verify the number of optimal number of clusters, a silhouette score plot was generated to compare the silhouette coefficient. The ideal number of clusters across the plot is that which is close to +1 as it demonstrates samples being further away from the decision boundary (Pedregosa et al., 2011). Based on figure 5.25, it confirms the optimal number of clusters as 3 since it registers a value of 0.65, the highest across the region. With this, the clusters generated earlier are maintained. The code snippet for this evaluation is provided in appendix C.8.

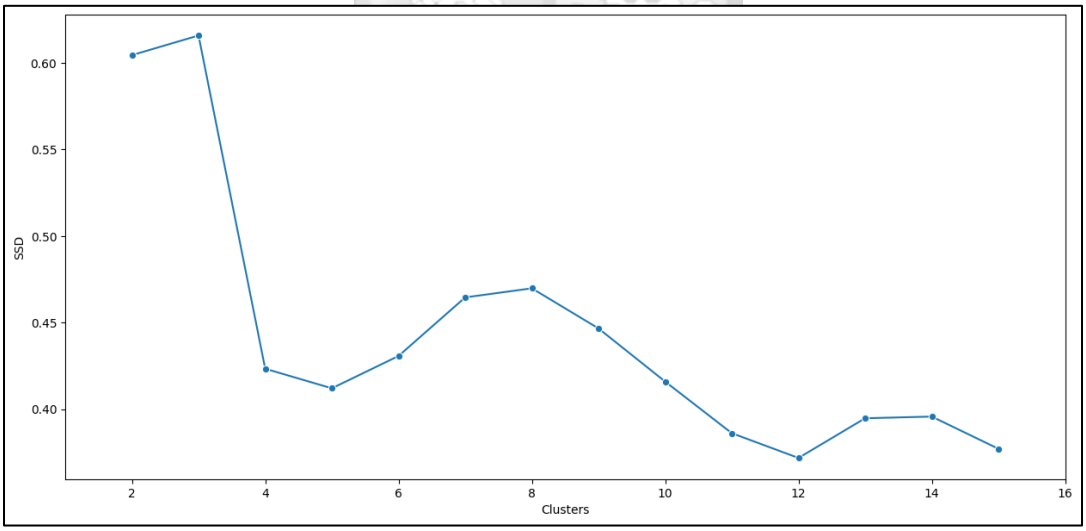


Figure 5.25: Silhouette Score for K-means Clustering.

5.5.3. Food Prediction

A decision tree classifier was used in predicting categories of foods that can be supported by soil at a particular location. It takes in the soil nutrient value as input, assigns it to a category of foods with similar nutrient values. The category of foods is then returned as output and queried to the PostgresSQL database to provide foods belonging to the same category.

Two classifiers were developed, one to classify foods based on soil potassium and the other based on soil magnesium. Both models were trained on 70% of the clustered vegetable data and tested with the 30% remainder. The classifier for potassium achieved an accuracy score of 81.82% in prediction while the magnesium-based classifier achieved a score of 90.91%. This is presented in figure 5.26.

DTree Classifiers

```
# split foods into potassium, magnesium train and test
X_k, y_k = foods.iloc[:, [3]], foods.iloc[:, 4]
X_mg, y_mg = foods.iloc[:, [2]], foods.iloc[:, 4]
X_ktrain, X_ktest, y_ktrain, y_ktest = train_test_split(X_k, y_k, test_size=0.3, random_state=42)
X_mgtrain, X_mgtest, y_mgtrain, y_mgtest = train_test_split(X_mg, y_mg, test_size=0.3, random_state=42)
print(X_ktrain.shape)
print(X_ktest.shape)
print(y_mgtrain.shape)
print(y_mgtest.shape)
```

```
(25, 1)
(11, 1)
(25,)
(11,)
```

```
# Dtree modelling
k_dtree = dt.DecisionTreeClassifier(criterion = 'gini')
k_dtree.fit(X_ktrain, y_ktrain)
yk_pred = k_dtree.predict(X_ktest)
print("K-tree Accuracy: ",accuracy_score(y_ktest, yk_pred))

mg_dtree = dt.DecisionTreeClassifier(criterion = 'gini')
mg_dtree.fit(X_mgtrain, y_mgtrain)
ymg_pred = mg_dtree.predict(X_mgtest)

print("Mg-tree Accuracy: ",accuracy_score(y_mgtest, ymg_pred))
```

```
K-tree Accuracy:  0.8181818181818182
Mg-tree Accuracy:  0.9090909090909091
```

Figure 5.26: Soil Nutrient Classification to Food Categories

Chapter 6: Discussion

6.1. Overview

This chapter provides a summary of findings of the study based on the identified objectives in section 1.3.2. Moreover, it provides key insights gained from each objective that have largely contributed to the achievement of the general objective of developing a predictive model that determines vegetable derived macronutrients using georeferenced maps of soil macronutrients to meet the micronutrient requirements in diabetic patients.

6.2. Review of Research Objectives

6.2.1. Relationship between Soil Nutrients, Plant Nutrients and Human Nutrition

The relationship between soil macronutrients and plant macronutrients, and the consequent influence on human nutrition was investigated, and is inherently highly complex. Soils can undergo various chemical and biological processes over time such as weathering from parent material, and climatic conditions that can affect the nutritional content. However, these are less likely to have great impact on soil nutrition, compared to management of agricultural land where activities such as land abandonment and urban development (Cerdà et al., 2019; El-Ramady et al., 2022). Fertilisers, irrigation and other sustainable agricultural management practices at small scale can extend the productivity of soils over time (Silver et al., 2021), and this has enhanced food security in small scale agricultural zones in Kenya (Mburu et al., 2016; Wawire et al., 2021).

Nutrient intake by plants is also crucial to understanding the link to human health, as it directly affects the crop quality and eventual yield. El-Ramady et al. (2022) supports this notion and suggests the enhancement of agricultural practices to meet the nutritional needs of the growing population. Regarding nutrient presence in vegetables, it was observed that potassium content is more prevalent than magnesium as depicted in figure 5.17. This implies that vegetables can be prescribed as a rich source of potassium to control glucose among type-II diabetic patients (Stone et al., 2016).

6.2.2. Modelling Techniques applied in Digital Soil Mapping

The third research objective was discussed in extensive literature review, covering different aspects of DSM: georeferenced soil data, a set of environmental covariates and a predictive model to capture unseen sampled locations, as described in figure 2.2.

Firstly, the set of georeferenced soil data used in this study is legacy soil data, particularly the Africa Soil Profiles Database that consists of legacy soil profiles prepared between 2007 and 2012 for soil profiles collected in Kenya. Despite the view of such information being dated, Hendriks et al. (2019) highlights its importance to gain further understanding about the soil in an area e.g. the formation process, composition and structure. Moreover, recent studies (Hengl et al., 2021; Hengl, Leenaars, et al., 2017; Leenaars et al., 2018), have effectively applied this data and recommend for its usage to be extended in other socio-economic and health domains to understand the role of soil nutrients within these spheres (Hengl, Leenaars, et al., 2017).

Regarding the environmental covariates, a set of 14 covariates were used in developing the maps used in this study. With reference to the scorpan model (equation 2.2), climate and relief covariates are popularly applied in prediction of soil nutrients (Gao et al., 2021; Suleymanov et al., 2021). The results of the correlation between covariates and soil nutrients demonstrated are comparatively low as which is similar to the results in Gao et al. (2021), Suleymanov et al. (2021) and John et al., (2022). Covariates used in this study were applied at 250m resolution resulting in soil maps with pixels representing areas of 62500m^2 or roughly 15.44 acres of ground area. Much as this may not be representative of smallholder farms that cover less than a hectare (Minai et al., 2021), it provides a fair representation of soil nutrient values across a large geographic region.

Lastly, four techniques for modelling soil maps were extensively reviewed: ordinary kriging (OK) and inverse distance weighting (IDW) as geostatistical methods, support vector machines (SVM) as a machine learning-based technique and (multiple linear) regression kriging (MLRK) as a hybrid technique. The last incorporates multiple linear regression (MLR) as a machine learning method and kriging as a geostatistical technique to interpolate residual values from regression analysis. From studies applying these methods summarised in table 2.2, there was need to compare the performance of a hybrid geospatial model to those by machine learning and geostatistics, especially when applying legacy soil data.

6.2.3. Application of Hybrid Techniques in Generating Digital Soil Maps

To achieve the fourth research objective, four models based on four techniques – SVM, OK, MLR and MLRK were developed, and prediction performances evaluated. As depicted in table 5.5, SVM and OK resulted in the least RMSE value for exchangeable potassium (ExK) and exchangeable magnesium (ExMg), whereas the MLR and MLRK based models registered high levels of RMSE. These results can

be compared to those achieved in Suleymanov et al. (2021), Gao et al. (2021) and Minai et al. (2021), as demonstrated in table 2.2. Hengl et al. (2017) attributes this inconsistency by kriging-based models to sampling points being rather far apart which would not have much effect on output predictions.

Wadoux et al. (2021) demonstrates the issues around applying standard and spatial cross-validation techniques as a method for model training and testing and concludes they both result in biased estimates of map accuracies. It is recommended that standard sampling techniques for evaluating map accuracy be practiced and would result in valid maps. In this study, a random split of 80:20 training and test sets were derived to develop and test the models.

6.2.4. Testing of Developed Soil Maps in Determining Vegetable-Derived Macronutrients

To respond to the fifth research objective, the generated soil maps were applied to determine suitable vegetables that can be supported by soils at a specific location. This was achieved by use of a k-means clustering algorithm to categorise vegetables by their nutrient values, and a decision tree classifier to categorise incoming soil nutrient value to the suitable vegetable category. This approach is similar to that in (W. Chen et al., 2017) and (Umam et al., 2022). Results for classification were relatively high, with 81.82% accuracy registered for classification based on potassium content and 90.91% accuracy for classification by magnesium content.

VT OMNES VNVM SINT

Chapter 7: Conclusion and Recommendations

7.1. Introduction

This chapter provides concluding remarks to the study conducted and its findings. It also provides recommendations for improvement on the developed system as well as before its implementation in a real-world setting. Lastly, it also highlights several avenues that can be taken to expand the study within the research domain.

7.2. Conclusion

The study set out to develop a predictive model for determining vegetable derived macronutrients for a diabetic patient using a georeferenced map of soil macronutrients. To achieve this objective, this involved a review into the role of soil nutrients and its connection to plant nutrition and human health, particularly in the management of type-II diabetes. It also included an extensive review into digital soil mapping (DSM), a predictive modelling technique been greatly applied in development of maps that depict soil attributes. In DSM, the study highlighted hybrid modelling techniques, which involve a combination of a machine learning model to predict values at unseen locations from sampled locations, and a geostatistical model to capture spatial structure information of residual values that result from regression analysis. Lastly, the study also achieved its primary goal by testing the validity of soil maps in determining suitable vegetables that can be consumed by type-II diabetic patients.

Soil is often overlooked in discussions around management of patients of type-II diabetes. However, this study has successfully demonstrated that patients can effectively gain the needed nutritional requirements through plant-based foods that are directly supported by soil. It is highlighted, nevertheless, that the connection between soil, plant, and human health is intrinsically complex due to varying interconnected factors that may be difficult to capture. DSM remains a highly suitable technique to discover soil patterns within regions and legacy soil data still plays a key role in the generation of soil maps. In Kenya, effort is greatly needed to avail legacy soil information. This will factor for regeneration of relatively more accurate maps using advanced techniques, as well as allow for new forms of analyses that display the role of soil in related domains such as climate change, agriculture, and demography.

Modelling techniques applied in DSM were studied, with the highlight on hybrid modelling techniques – a combination of machine learning based models and geostatistical based models. The study included

one geostatistical model based on ordinary kriging (OK), two based on machine learning models – support vector machines (SVM) and multiple linear regression (MLR) and one hybrid technique based on multiple linear regression kriging (MLRK). This involved a combination of multiple linear regression to predict nutrient values at unseen locations based on the recorded observations and a set of environmental covariates and model the residual values from regression by kriging. It was expected that the MLRK model would outperform the singular geostatistical and machine learning models, but the SVM method proved to produce minimal errors compared to other methods for both exchangeable potassium and exchangeable magnesium. This demonstrates machine learning techniques are a sufficient mapping technique specifically for relatively large extents.

Lastly, a predictive model was developed to test for the usability of the soil maps in determining suitable plant-based foods that would meet both patient nutritional requirements; and whose crop will be supported by the soil in a specific location. This study focused comparing foods based on potassium content and magnesium content and achieved suitable levels of accuracy in determination. Nutritionists applying this tool can, therefore, provide location-based food recommendations to type-II diabetic patients based on soil nutrient values.

7.3. Recommendations

Before deployment of this tool, specialists in nutrition and soil science should further verify the applicability of the developed maps and food determinant tool. This is due to their role in guiding patient health and nutrition, which demands for high levels of accuracy. The tool could be expanded to include other nutritional components that are crucial to diabetic patients; and would have either direct or indirect relationships to soil such as meats, dairy products, and fibre. External factors such as environmental conditions, crop yield, demographics (for instance income, household size and culture) could potentially affect recommendations made for patient health, and thus can be investigated in future studies. Other modules for management of diabetic patients such as physical activity monitoring and medication management can also be captured to cover different facets of patient health.

To enhance soil mapping and analysis in Kenya, the expansion and availability of more legacy soil information in the public domain would be beneficial for future studies. For instance, time series analyses can be conducted to demonstrate changes in soil components and related factors over time.

References

Asrat, T. G., Sakrabani, R., Corstanje, R., Breure, T., Hassall, K. L., Kebede, F., & Haefele, S. M. (2023). Spectral soil analysis for fertilizer recommendations by coupling with QUEFTS for maize in East Africa: A sensitivity analysis. *Geoderma*, *432*, 116397. https://doi.org/10.1016/j.geoderma.2023.116397

Babak, O., & Deutsch, C. V. (2009). Statistical approach to inverse distance interpolation. *Stochastic Environmental Research and Risk Assessment*, *23*(5), 543–553. https://doi.org/10.1007/s00477-008-0226-6

Baker, J., White, N., & Mengersen, K. (2014). Missing in space: An evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. *International Journal of Health Geographics*, *13*(1), 47. https://doi.org/10.1186/1476-072X-13-47

Baqar, S., Michalopoulos, A., Jerums, G., & Ekinci, E. I. (2020). Dietary sodium and potassium intake in people with diabetes: Are guidelines being met? *Nutrition & Diabetes*, *10*(1), Article 1. https://doi.org/10.1038/s41387-020-0126-5

Barrena-González, J., Lavado Contador, J. F., & Pulido Fernández, M. (2022). Mapping Soil Properties at a Regional Scale: Assessing Deterministic vs. Geostatistical Interpolation Methods at Different Soil Depths. *Sustainability*, *14*(16), Article 16. https://doi.org/10.3390/su141610049

Behrens, T., Zhu, A.-X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, *155*(3), 175–185. https://doi.org/10.1016/j.geoderma.2009.07.010

Berkhout, E. D., Malan, M., & Kram, T. (2019). Better soils for healthier lives? An econometric assessment of the link between soil nutrients and malnutrition in Sub-Saharan Africa. *PLoS ONE*, *14*(1), e0210642. https://doi.org/10.1371/journal.pone.0210642

Bolat, F., Bulut, S., Günlü, A., Ercanlı, İ., & Şenyurt, M. (2020). Regression kriging to improve basal area and growing stock volume estimation based on remotely sensed data, terrain indices and forest inventory of black pine forests. *New Zealand Journal of Forestry Science*, *50*. https://doi.org/10.33494/nzjfs502020x49x

Bowen, C. K., Schuman, G. E., Olson, R. A., & Ingram, J. (2005). Influence of Topsoil Depth on Plant and Soil Attributes of 24-Year Old Reclaimed Mined Lands. *Arid Land Research and Management*, *19*(3), 267–284. https://doi.org/10.1080/15324980590951441

Brevik, E. C., Slaughter, L., Singh, B. R., Steffan, J. J., Collier, D., Barnhart, P., & Pereira, P. (2020). Soil and Human Health: Current Status and Future Needs. *Air, Soil and Water Research*, *13*, 1178622120934441. https://doi.org/10.1177/1178622120934441

Brevik, E. C., Steffan, J. J., Burgess, L. C., & Cerdà, A. (2017). Links Between Soil Security and the Influence of Soil on Human Health. In D. J. Field, C. L. S. Morgan, & A. B. McBratney (Eds.), *Global Soil Security* (pp. 261–274). Springer International Publishing. https://doi.org/10.1007/978-3-319-43394-3_24

Brindha, K., Taie Semiromi, M., Boumaiza, L., & Mukherjee, S. (2023). Comparing Deterministic and Stochastic Methods in Geospatial Analysis of Groundwater Fluoride Concentration. *Water*, *15*(9), Article 9. https://doi.org/10.3390/w15091707

Buenemann, M., Coetzee, M. E., Kutuahupira, J., Maynard, J. J., & Herrick, J. E. (2023). Errors in soil maps: The need for better on-site estimates and soil map predictions. *PLOS ONE*, *18*(1), e0270176. https://doi.org/10.1371/journal.pone.0270176

Cerdà, A., Ackermann, O., Terol, E., & Rodrigo-Comino, J. (2019). Impact of Farmland Abandonment on Water Resources and Soil Conservation in Citrus Plantations in Eastern Spain. *Water*, *11*(4), Article 4. https://doi.org/10.3390/w11040824

Chagumaira, C., Chimungu, J. G., Nalivata, P. C., Broadley, M. R., Nussbaum, M., Milne, A. E., & Lark, R. M. (2022). Mapping soil micronutrient concentration at national-scale: An illustration of a decision process framework. *EGUsphere*, 1–33. https://doi.org/10.5194/egusphere-2022-583

Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., & Walter, C. (2022). Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. *Geoderma*, *409*, 115567. https://doi.org/10.1016/j.geoderma.2021.115567

Chen, W., Chen, S., Zhang, H., & Wu, T. (2017). A hybrid prediction model for type 2 diabetes using K-means and decision tree. *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 386–390. https://doi.org/10.1109/ICSESS.2017.8342938

Comerford, N. B. (2005). Soil Factors Affecting Nutrient Bioavailability. In H. BassiriRad (Ed.), *Nutrient Acquisition by Plants: An Ecological Perspective* (pp. 1–14). Springer. https://doi.org/10.1007/3-540-27675-0_1

Dansinger, M. (2023, March 18). *The Link Between Salt and Diabetes*. WebMD. https://www.webmd.com/diabetes/diabetes-understanding-salt

Dasgupta, S., Debnath, S., Das, A., Biswas, A., Weindorf, D. C., Li, B., Kumar Shukla, A., Das, S., Saha, S., & Chakraborty, S. (2023). Developing regional soil micronutrient management strategies through ensemble learning based digital soil mapping. *Geoderma*, *433*, 116457. https://doi.org/10.1016/j.geoderma.2023.116457

Dennis, A., Wixom, B. H., & Roth, R. M. (2012). *Systems Analysis and Design* (5th ed.). Wiley Publishing.

Dias, J. S. (2019). Nutritional Quality and Effect on Disease Prevention of Vegetables. *Food and Nutrition Sciences*, *10*(4), Article 4. https://doi.org/10.4236/fns.2019.104029

Dubey, P., Thakur, V., & Chattopadhyay, M. (2020). Role of Minerals and Trace Elements in Diabetes and Insulin Resistance. *Nutrients*, *12*(6), 1864. https://doi.org/10.3390/nu12061864

El-Ramady, H., Hajdú, P., Törős, G., Badgar, K., Llanaj, X., Kiss, A., Abdalla, N., Omara, A. E.-D., Elsakhawy, T., Elbasiouny, H., Elbehiry, F., Amer, M., El-Mahrouk, M. E., & Prokisch, J. (2022). Plant Nutrition for Human Health: A Pictorial Review on Plant Bioactive Compounds for Sustainable Agriculture. *Sustainability*, *14*(14), Article 14. https://doi.org/10.3390/su14148329

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020). Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. *Remote Sensing*, *12*(14), Article 14. https://doi.org/10.3390/rs12142234

Fageria, N. K., Gheyi, H. R., & Moreira, A. (2011). Nutrient Bioavailability in Salt Affected Soils. *Journal of Plant Nutrition*, *34*(7), 945–962. https://doi.org/10.1080/01904167.2011.555578

FAO. (2018). *Soil Organic Carbon Mapping Cookbook: 2nd edition* (Y. Yigini, G.F. Olmedo, S. Reiter, R. Baritz, K. Viatkin, & R. R. Vargas, Eds.; 2nd ed.). Food & Agriculture Org.

FAO. (2022). *Soils for nutrition: State of the art*. FAO. https://doi.org/10.4060/cc0900en

FAO & Government of Kenya. (2018). *Kenya Food Composition Tables* [dataset]. https://www.fao.org/3/I8897EN/i8897en.pdf

Fernández, F. G., & Hoeft, R. G. (2009). Managing soil pH and crop nutrients. In *Illinois Agronomy Handbook* (Vol. 24, pp. 91–112). University of Illinois at Urbana-Champaign, College of Agriculture, Cooperative Extension Service.

Folorunso, O., Ojo, O., Busari, M., Adebayo, M., Joshua, A., Folorunso, D., Ugwunna, C. O., Olabanjo, O., & Olabanjo, O. (2023). Exploring Machine Learning Models for Soil Nutrient Properties Prediction: A Systematic Review. *Big Data and Cognitive Computing*, *7*(2), Article 2. https://doi.org/10.3390/bdcc7020113

Gao, L., Huang, M., Zhang, W., Qiao, L., Wang, G., & Zhang, X. (2021). Comparative Study on Spatial Digital Mapping Methods of Soil Nutrients Based on Different Geospatial Technologies. *Sustainability*, *13*(6), Article 6. https://doi.org/10.3390/su13063270

Gomaa, H. (2011). *Software Modeling and Design: UML, Use Cases, Patterns, and Software Architectures*. Cambridge University Press.

Gray, A., & Threlkeld, R. J. (2000). Nutritional Recommendations for Individuals with Diabetes. In K. R. Feingold, B. Anawalt, M. R. Blackman, A. Boyce, G. Chrousos, E. Corpas, W. W. de Herder, K. Dhatariya, K. Dungan, J. Hofland, S. Kalra, G. Kaltsas, N. Kapoor, C. Koch, P. Kopp, M. Korbonits, C. S. Kovacs, W. Kuohung, B. Laferrère, … D. P. Wilson (Eds.), *Endotext*. MDText.com, Inc. http://www.ncbi.nlm.nih.gov/books/NBK279012/

Hanrahan, G., Zhu, J., Gibani, S., & Patil, D. G. (2005). CHEMOMETRICS AND STATISTICS | Experimental Design. In P. Worsfold, A. Townshend, & C. Poole (Eds.), *Encyclopedia of Analytical Science (Second Edition)* (pp. 8–13). Elsevier. https://doi.org/10.1016/B0-12-369397-7/00079-0

Hendriks, C. M. J., Stoorvogel, J. J., Lutz, F., & Claessens, L. (2019). When can legacy soil data be used, and when should new data be collected instead? *Geoderma*, *348*, 181–188. https://doi.org/10.1016/j.geoderma.2019.04.026

Hengl, T., Jesus, J. M. de, Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hengl, T., Leenaars, J. G. B., Shepherd, K. D., Walsh, M. G., Heuvelink, G. B. M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I., & Kwabena, N. A. (2017). Soil nutrient maps of Sub-Saharan Africa: Assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*, *109*(1), 77–102. https://doi.org/10.1007/s10705-017-9870-x

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F. B. T., … Crouch, J. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using

two-scale ensemble machine learning. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-85639-y

Heung, B., Saurette, D., & Bulmer, C. E. (2021). Digital Soil Mapping. In Krzic, M., Walley, F.L., Diochon, A., Paré, M.C., & Farrell, R.E. (Eds.), *Digging into Canadian soils: An Introduction to Soil Science*. Canadian Society of Soil Science. https://openpress.usask.ca/soilscience/chapter/digital-soil-mapping/

Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., Thornton, P. L., & Haire-Joshu, D. (2020). Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care*, *44*(1), 258–279. https://doi.org/10.2337/dci20-0053

Howell, C. R., Su, W., Nassel, A. F., Agne, A. A., & Cherrington, A. L. (2020). Area based stratified random sampling using geospatial technology in a community-based survey. *BMC Public Health*, *20*(1), 1678. https://doi.org/10.1186/s12889-020-09793-0

Huang, S., Wang, P., Yamaji, N., & Ma, J. F. (2020). Plant Nutrition for Human Nutrition: Hints from Rice Research and Future Perspectives. *Molecular Plant*, *13*(6), 825–835. https://doi.org/10.1016/j.molp.2020.05.007

IBM Corporation. (2021, August 17). *CRISP-DM Help Overview*. CRISP-DM Help Overview - IBM Documentation. https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

IEBC. (2018). *Kenya—Subnational Administrative Boundaries* [Shapefiles]. Information Technology Outreach Services (ITOS). https://data.humdata.org/dataset/cod-ab-ken

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Jeon, J., Jang, J., & Park, K. (2018). Effects of Consuming Calcium-Rich Foods on the Incidence of Type 2 Diabetes Mellitus. *Nutrients*, *11*(1), 31. https://doi.org/10.3390/nu11010031

John, K., Bouslihim, Y., Isong, I. A., Hssaini, L., Razouk, R., Kebonye, N. M., Agyeman, P. C., Penížek, V., & Zádorová, T. (2022). Mapping soil nutrients via different covariates combinations: Theory and an example from Morocco. *Ecological Processes*, *11*(1), 23. https://doi.org/10.1186/s13717-022-00368-y

Johnson, V. J., & Mirza, A. (2020). Role of Macro and Micronutrients in the Growth and Development of Plants. *International Journal of Current Microbiology and Applied Sciences*, *9*(11), 576–587. https://doi.org/10.20546/ijcmas.2020.911.071

Karega, L. A. (2022). *A Location-aware nutritional needs prediction tool for type II Diabetic patients: Case Kenya* [Strathmore University]. http://hdl.handle.net/11071/13208

Keesstra, S. D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J. N., Pachepsky, Y., van der Putten, W. H., Bardgett, R. D., Moolenaar, S., Mol, G., Jansen, B., & Fresco, L. O. (2016). The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. *SOIL*, *2*(2), 111–128. https://doi.org/10.5194/soil-2-111-2016

Kendall, K. E., & Kendall, J. E. (2011). *Systems Analysis and Design* (8th ed). Pearson Prentice Hall.

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, *81*, 401–418. https://doi.org/10.1016/j.apm.2019.12.016

Kienast-Brown, S., Libohova, Z., & Boettinger, J. (2017). Digital Soil Mapping. In C. Ditzler, K. Scheffe, & H. C. Monger (Eds.), *Soil survey manual* (USDA Handbook 18, pp. 295–354). Government Printing Office. https://www.nrcs.usda.gov/sites/default/files/2022-09/The-Soil-Survey-Manual.pdf

Kimsey, M. J., Laing, L. E., Anderson, S. M., Bruggink, J., Campbell, S., Diamond, D., Domke, G. M., Gries, J., Holub, S. M., Nowacki, G., Page-Dumroese, D. S., Perry, C. H. (Hobie), Rustad, L. E., Stephens, K., & Vaughan, R. (2020). Soil Mapping, Monitoring, and Assessment. In R. V. Pouyat, D. S. Page-Dumroese, T. Patel-Weynand, & L. H. Geiser (Eds.), *Forest and Rangeland Soils of the United States Under Changing Conditions: A Comprehensive Science Synthesis* (pp. 169–188). Springer International Publishing. https://doi.org/10.1007/978-3-030-45216-2_9

Kiprotich, D., Chege, P., & Mituki, D. (2022). Dietary Practices of Patients with Type 2 Diabetes Attending Clinic at Nakuru Level 6 Hospital, Kenya. *African Journal of Nutrition and Dietetics*, *1*(1), Article 1. https://doi.org/10.58460/ajnd.v1i1.4

Kotu, V., & Deshpande, B. (2019). Chapter 4—Classification. In V. Kotu & B. Deshpande (Eds.), *Data Science (Second Edition)* (pp. 65–163). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-814761-0.00004-6

Kumar, N., & Sinha, N. K. (2018). Geostatistics: Principles and Applications in Spatial Mapping of Soil Properties. In G. P. O. Reddy & S. K. Singh (Eds.), *Geospatial Technologies in Land Resources Mapping, Monitoring and Management* (pp. 143–159). Springer International Publishing. https://doi.org/10.1007/978-3-319-78711-4_8

Lacoste, M., Mulder, V. L., Richer-de-Forges, A. C., Martin, M. P., & Arrouays, D. (2016). Evaluating large-extent spatial modeling approaches: A case study for soil depth for France. *Geoderma Regional*, *7*(2), 137–152. https://doi.org/10.1016/j.geodrs.2016.02.006

Lagacherie, P. (2008). Digital Soil Mapping: A State of the Art. In A. E. Hartemink, A. McBratney, & M. de L. Mendonça-Santos (Eds.), *Digital Soil Mapping with Limited Data* (pp. 3–14). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8592-5_1

Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma*, *352*, 395–413. https://doi.org/10.1016/j.geoderma.2019.05.031

Leenaars, J. G. B., Claessens, L., Heuvelink, G. B. M., Hengl, T., Ruiperez González, M., van Bussel, L. G. J., Guilpart, N., Yang, H., & Cassman, K. G. (2018). Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa. *Geoderma*, *324*, 18–36. https://doi.org/10.1016/j.geoderma.2018.02.046

Leenaars, J. G. B., Oostrum, van A., & Gonzalez, M. R. (2014). *Africa Soil Profiles Database* (ISRIC Report 2014/01, p. 162). Africa Soil Information Service (AfSIS) project and ISRIC - World Soil Information. https://data.isric.org/geonetwork/srv/api/records/b88870b4-6af8-4e78-a3ac-38871d757525

Li, H., Leng, W., Zhou, Y., Chen, F., Xiu, Z., & Yang, D. (2014). Evaluation Models for Soil Nutrient Based on Support Vector Machine and Artificial Neural Networks. *The Scientific World Journal*, *2014*, e478569. https://doi.org/10.1155/2014/478569

Lu, Y., Liu, F., Zhao, Y., Song, X., & Zhang, G. (2019). An integrated method of selecting environmental covariates for predictive soil depth mapping. *Journal of Integrative Agriculture*, *18*(2), 301–315. https://doi.org/10.1016/S2095-3119(18)61936-7

Manyara, A. M., Mwaniki, E., Gill, J. M. R., & Gray, C. M. (2024). Perceptions of diabetes risk and prevention in Nairobi, Kenya: A qualitative and theory of change development study. *PLOS ONE*, *19*(2), e0297779. https://doi.org/10.1371/journal.pone.0297779

Matinfar, H. R., Maghsodi, Z., Mousavi, S. R., & Rahmani, A. (2021). Evaluation and Prediction of Topsoil organic carbon using Machine learning and hybrid models at a Field-scale. *CATENA*, *202*, 105258. https://doi.org/10.1016/j.catena.2021.105258

Mburu, S. W., Koskey, G., Kimiti, J. M., Ombori, O., Maingi, J. M., & Njeru, E. M. (2016). Agrobiodiversity conservation enhances food security in subsistence-based farming systems of

Eastern Kenya. *Agriculture & Food Security*, *5*(1), 19. https://doi.org/10.1186/s40066-016-0068-2

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

McBratney, A., Field, D. J., & Koch, A. (2014). The dimensions of soil security. *Geoderma*, *213*, 203–213. https://doi.org/10.1016/j.geoderma.2013.08.013

Minai, J. O., Libohova, Z., & Schulze, D. G. (2021). Spatial prediction of soil properties for the Busia area, Kenya using legacy soil data. *Geoderma Regional*, *25*, e00366. https://doi.org/10.1016/j.geodrs.2021.e00366

Minari, T. P., Tácito, L. H. B., Yugar, L. B. T., Ferreira-Melo, S. E., Manzano, C. F., Pires, A. C., Moreno, H., Vilela-Martin, J. F., Cosenso-Martin, L. N., & Yugar-Toledo, J. C. (2023). Nutritional Strategies for the Management of Type 2 Diabetes Mellitus: A Narrative Review. *Nutrients*, *15*(24), Article 24. https://doi.org/10.3390/nu15245096

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

Mishra, U., Gautam, S., Riley, W. J., & Hoffman, F. M. (2020). Ensemble Machine Learning Approach Improves Predicted Spatial Variation of Surface Soil Organic Carbon Stocks in Data-Limited Northern Circumpolar Region. *Frontiers in Big Data*, *3*. https://www.frontiersin.org/articles/10.3389/fdata.2020.528441

Ngunjiri, M. W., Libohova, Z., Minai, J. O., Serrem, C., Owens, P. R., & Schulze, D. G. (2019). Predicting soil types and soil properties with limited data in the Uasin Gishu Plateau, Kenya. *Geoderma Regional*, *16*, e00210. https://doi.org/10.1016/j.geodrs.2019.e00210

Nguyen, H. D., Oh, H., & Kim, M.-S. (2022). Higher intakes of nutrients are linked with a lower risk of cardiovascular diseases, type 2 diabetes mellitus, arthritis, and depression among Korean adults. *Nutrition Research*, *100*, 19–32. https://doi.org/10.1016/j.nutres.2021.11.003

Nikparvar, B., & Thill, J.-C. (2021). Machine Learning of Spatial Data. *ISPRS International Journal of Geo-Information*, *10*(9), Article 9. https://doi.org/10.3390/ijgi10090600

Nutritionists and Dieticians Act, Pub. L. No. 18 of 2007, 19 (2008). https://www.kndi.institute/pages/downloads/NutritionistsAct.pdf

Onteri, S. N., Kariuki, J., Mathu, D., Wangui, A. M., Magige, L., Mutai, J., Chuchu, V., Karanja, S., Ahmed, I., Mokua, S., Otambo, P., & Bukania, Z. (2023). Diabetes health care specific services

readiness and availability in Kenya: Implications for Universal Health Coverage. *PLOS Global Public Health*, *3*(9), e0002292. https://doi.org/10.1371/journal.pgph.0002292

Ouabo, R. E., Sangodoyin, A. Y., & Ogundiran, M. B. (2020). Assessment of Ordinary Kriging and Inverse Distance Weighting Methods for Modeling Chromium and Cadmium Soil Pollution in E-Waste Sites in Douala, Cameroon. *Journal of Health & Pollution*, *10*(26), 200605. https://doi.org/10.5696/2156-9614-10.26.200605

Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *SOIL*, *6*(1), 35–52. https://doi.org/10.5194/soil-6-35-2020

Patriche, C. V., Roşca, B., Pîrnău, R. G., & Vasiliniuc, I. (2023). Spatial modelling of topsoil properties in Romania using geostatistical methods and machine learning. *PLOS ONE*, *18*(8), e0289286. https://doi.org/10.1371/journal.pone.0289286

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Piedallu, C., Pedersoli, E., Chaste, E., Morneau, F., Seynave, I., & Gégout, J.-C. (2022). Optimal resolution of soil properties maps varies according to their geographical extent and location. *Geoderma*, *412*, 115723. https://doi.org/10.1016/j.geoderma.2022.115723

Pierangeli, L. M. P., Silva, S. H. G., Teixeira, A. F. dos S., Mancini, M., Andrade, R., de Menezes, M. D., Marques, J. J., Weindorf, D. C., & Curi, N. (2022). Combining Proximal and Remote Sensors in Spatial Prediction of Five Micronutrients and Soil Texture in a Case Study at Farmland Scale in Southeastern Brazil. *Agronomy*, *12*(11), Article 11. https://doi.org/10.3390/agronomy12112699

Raynaud, X., & Leadley, P. W. (2004). Soil Characteristics Play a Key Role in Modeling Nutrient Competition in Plant Communities. *Ecology*, *85*(8), 2200–2214. https://doi.org/10.1890/03-0817

Rekik, F., & van Es, H. M. (2022). Soils and Human Health: Connections Between Geo-Environmental, Socio-Demographic, and Lifestyle factors and Nutrition of Tribal Women of Jharkhand, India. *Frontiers in Soil Science*, *2*. https://www.frontiersin.org/articles/10.3389/fsoil.2022.901843

Ruthsatz, M., & Candeias, V. (2020). Non-communicable disease prevention, nutrition and aging. *Acta Bio Medica : Atenei Parmensis*, *91*(2), 379–388. https://doi.org/10.23750/abm.v91i2.9721

Sahu, B., Ghosh, A. K., & Seema. (2021). Deterministic and geostatistical models for predicting soil organic carbon in a 60 ha farm on Inceptisol in Varanasi, India. *Geoderma Regional*, *26*, e00413. https://doi.org/10.1016/j.geodrs.2021.e00413

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. *2021 IEEE International Conference on Big Data (Big Data)*, 2337–2344. https://doi.org/10.1109/BigData52589.2021.9671634

Sami, W., Ansari, T., Butt, N. S., & Hamid, M. R. A. (2017). Effect of diet on type 2 diabetes mellitus: A review. *International Journal of Health Sciences*, *11*(2), 65.

Sarah, K., Abdillahi, H. S., Reuben, M., & Lydia, A. (2021). Prevalence and risk factors associated with diabetes in Meru County, Kenya: A cross-sectional study. *International Journal of Diabetes in Developing Countries*, *41*(3), 412–418. https://doi.org/10.1007/s13410-020-00902-8

Schönfeldt, H. C., Pretorius, B., & Hall, N. (2016). Bioavailability of Nutrients. In B. Caballero, P. M. Finglas, & F. Toldrá (Eds.), *Encyclopedia of Food and Health* (pp. 401–406). Academic Press. https://doi.org/10.1016/B978-0-12-384947-2.00068-4

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Silver, W. L., Perez, T., Mayer, A., & Jones, A. R. (2021). The role of soil in the contribution of food and feed. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1834), 20200181. https://doi.org/10.1098/rstb.2020.0181

Sommerville, I. (2011). *Software Engineering* (9th ed). Pearson.

State of New South Wales through Local Land Services. (2020). *Cation Exchange Capacity—Fact Sheet 4* (p. 4). Local Land Services. https://www.lls.nsw.gov.au/__data/assets/pdf_file/0003/1270524/4-Cation-Exchange-Capacity_FINAL.pdf

Stewart, Z. P., Pierzynski, G. M., Middendorf, B. J., & Prasad, P. V. V. (2020). Approaches to improve soil fertility in sub-Saharan Africa. *Journal of Experimental Botany*, *71*(2), 632–641. https://doi.org/10.1093/jxb/erz446

Stone, M. S., Martyn, L., & Weaver, C. M. (2016). Potassium Intake, Bioavailability, Hypertension, and Glucose Control. *Nutrients*, *8*(7). https://doi.org/10.3390/nu8070444

Suleymanov, A., Abakumov, E., Suleymanov, R., Gabbasova, I., & Komissarov, M. (2021). The Soil Nutrient Digital Mapping for Precision Agriculture Cases in the Trans-Ural Steppe Zone of Russia Using Topographic Attributes. *ISPRS International Journal of Geo-Information*, *10*(4), Article 4. https://doi.org/10.3390/ijgi10040243

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., & Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, *183*, 109119. https://doi.org/10.1016/j.diabres.2021.109119

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1), 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Szerement, J., Szatanik-Kloc, A., Mokrzycki, J., & Mierzwa-Hersztek, M. (2022). Agronomic Biofortification with Se, Zn, and Fe: An Effective Strategy to Enhance Crop Nutritional Quality and Stress Defense—A Review. *Journal of Soil Science and Plant Nutrition*, *22*(1), 1129–1159. https://doi.org/10.1007/s42729-021-00719-2

Tarasov, D. A., Buevich, A. G., Sergeev, A. P., & Shichkin, A. V. (2018). High variation topsoil pollution forecasting in the Russian Subarctic: Using artificial neural networks combined with residual kriging. *Applied Geochemistry*, *88*, 188–197. https://doi.org/10.1016/j.apgeochem.2017.07.007

Tedre, M., & Moisseinen, N. (2014). Experiments in Computing: A Survey. *The Scientific World Journal*, *2014*, e549398. https://doi.org/10.1155/2014/549398

Umam, M. W. F., Fatekurohman, M., & Anggraeni, D. (2022). Hybrid clustering and classification methods to find out the pattern of the spread of covid-19 in East Java province. *Journal of Physics: Conference Series*, *2157*(1), 012030. https://doi.org/10.1088/1742-6596/2157/1/012030

Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692

Wadoux, A. M. J.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, *210*, 103359. https://doi.org/10.1016/j.earscirev.2020.103359

Watson, S. (2022, June 20). *The Link Between Diabetes and Potassium*. WebMD. https://www.webmd.com/diabetes/potassium-diabetes

Wawire, A. W., Csorba, Á., Tóth, J. A., Michéli, E., Szalai, M., Mutuma, E., & Kovács, E. (2021). Soil fertility management among smallholder farmers in Mount Kenya East region. *Heliyon*, *7*(3), e06488. https://doi.org/10.1016/j.heliyon.2021.e06488

Wekesah, F. M., Kyobutungi, C., Grobbee, D. E., & Klipstein-Grobusch, K. (2019). Understanding of and perceptions towards cardiovascular diseases and their risk factors: A qualitative study among residents of urban informal settings in Nairobi. *BMJ Open*, *9*(6), e026852. https://doi.org/10.1136/bmjopen-2018-026852

Xiaobo, Z., Jiewen, Z., Povey, M. J. W., Holmes, M., & Hanpin, M. (2010). Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, *667*(1–2), 14–32. https://doi.org/10.1016/j.aca.2010.03.048

Yasrebi, J., Saffari, M., Fathi, H., Karimian, N., Moazallahi, M., Gazni, R., & others. (2009). Evaluation and comparison of ordinary kriging and inverse distance weighting methods for prediction of spatial variability of some soil chemical parameters. *Research Journal of Biological Sciences*, *4*(1), 93–102.

Zelman, K. (2022, November 16). *Calcium and Vitamin D: Top Foods to Prevent Osteoporosis*. Calcium and Vitamin D: Top Foods to Prevent Osteoporosis. https://www.webmd.com/food-recipes/calcium-vitamin-d-foods

Zewide, I., & Melash, W. (2021). Review on Macronutrient in Agronomy Crops. *Nutrition and Food Processing*, *4*(6), 4. https://doi.org/10.31579/2637-8914/062

Appendix

Appendix A: Ethical Clearance Certification

Strathmore UNIVERSITY

5th February 2024

Mr Vikiru Allan,
allan.vikiru@strathmore.edu

Dear Mr Vikiru,

RE: A Predictive Model for Determining Vegetable-Derived Macronutrients for a Diabetic Patient

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC1998/23.** The approval period is from **5th February 2024 to 4th February 2025.**
This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

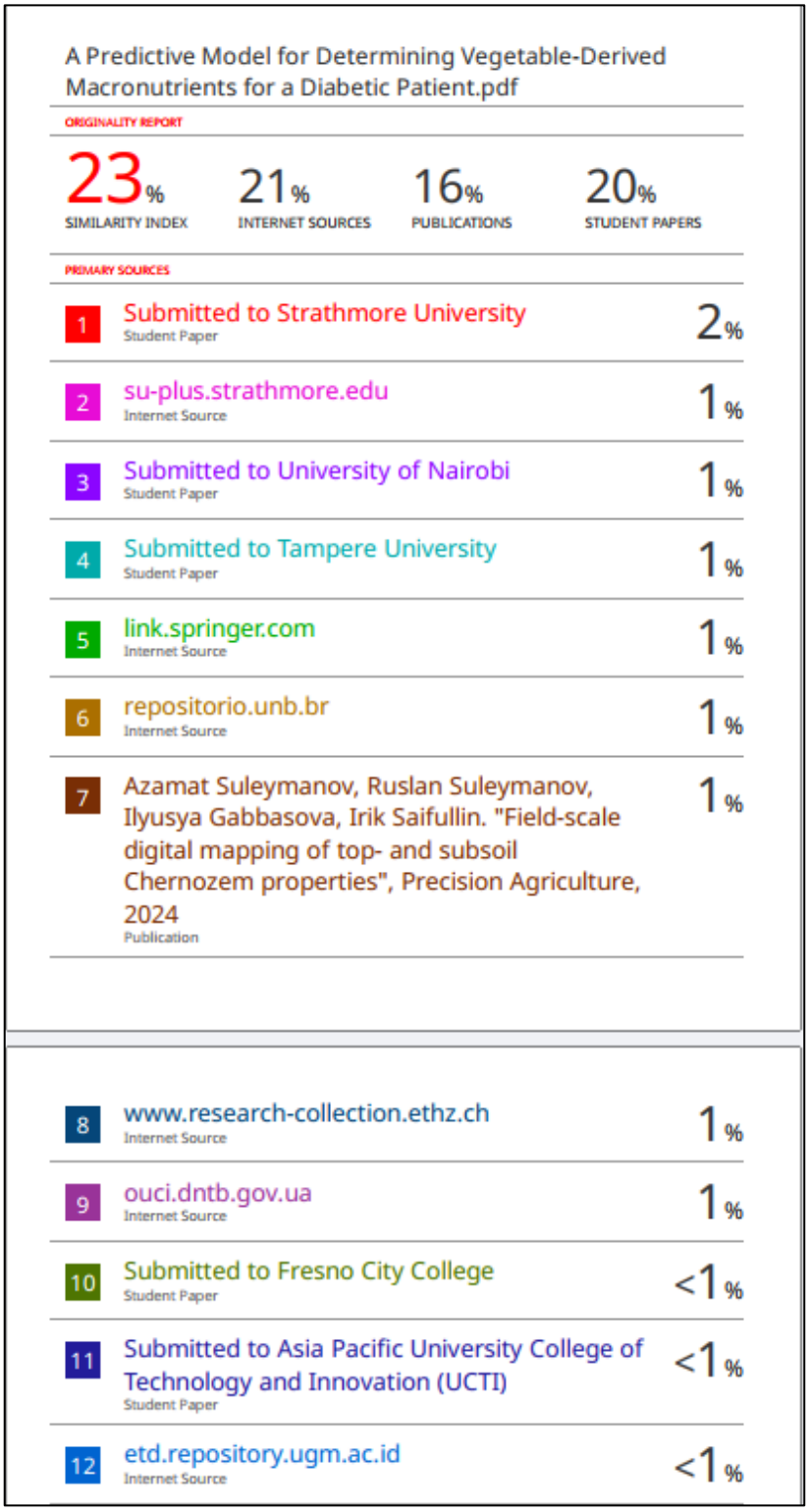
Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
05-Feb-2024
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix B: Originality Report



A Predictive Model for Determining Vegetable-Derived Macronutrients for a Diabetic Patient.pdf

ORIGINALITY REPORT

23%	21%	16%	20%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Strathmore University Student Paper	2%
2	su-plus.strathmore.edu Internet Source	1%
3	Submitted to University of Nairobi Student Paper	1%
4	Submitted to Tampere University Student Paper	1%
5	link.springer.com Internet Source	1%
6	repositorio.unb.br Internet Source	1%
7	Azamat Suleymanov, Ruslan Suleymanov, Ilyusya Gabbasova, Irik Saifullin. "Field-scale digital mapping of top- and subsoil Chernozem properties", Precision Agriculture, 2024 Publication	1%
8	www.research-collection.ethz.ch Internet Source	1%
9	ouci.dntb.gov.ua Internet Source	1%
10	Submitted to Fresno City College Student Paper	<1%
11	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
12	etd.repository.ugm.ac.id Internet Source	<1%

Appendix C: System Implementation Code Snippets

C.1. Points-in-Polygon Operation

```
QGIS version: 3.34.3-Prizren

QGIS code revision: 47373234ac

Qt version: 5.15.3

Python version: 3.9.18

GDAL version: 3.8.3

GEOS version: 3.12.1-CAPI-1.18.1

PROJ version: Rel. 9.3.1, December 1st, 2023

PDAL version: 2.6.0 (git-version: 3fced5)

Algorithm started at: 2024-03-18T22:06:33

Algorithm 'Count points in polygon' starting…

Input parameters:

{     'CLASSFIELD'     :     '',     'FIELD'     :     'num_sp',     'OUTPUT'     :
'D:/r/KenyaDSM/data/soil/ke_sp/pts_per_cty.shp',             'POINTS'             :
'delimitedtext://file:///D:/r/KenyaDSM/data/soil/ke_sp/ke_sp.csv?type=csv&maxFields
=10000&detectTypes=yes&xField=loc_longitude&yField=loc_latitude&crs=EPSG:4326&spati
alIndex=no&subsetIndex=no&watchFile=no',                 'POLYGONS'                 :
'D:/r/KenyaDSM/data/ke/ken_admbnda_adm1_iebc_20191031.shp', 'WEIGHT' : '' }


No spatial index exists for points layer, performance will be severely degraded

Execution completed in 1.21 seconds

Results:

{'OUTPUT': 'D:/r/KenyaDSM/data/soil/ke_sp/pts_per_cty.shp'}


Loading resulting layers

Algorithm 'Count points in polygon' finished
```

C.2. Extract by Location Function

```
QGIS version: 3.34.3-Prizren

QGIS code revision: 47373234ac

Qt version: 5.15.3

Python version: 3.9.18

GDAL version: 3.8.3

GEOS version: 3.12.1-CAPI-1.18.1

PROJ version: Rel. 9.3.1, December 1st, 2023

PDAL version: 2.6.0 (git-version: 3fced5)

Algorithm started at: 2024-03-18T22:26:14

Algorithm 'Extract by location' starting…

Input parameters:

{ 'INPUT' :
'delimitedtext://file:///D:/r/KenyaDSM/data/soil/ke_sp/ke_sp.csv?type=csv&maxFields
=10000&detectTypes=yes&xField=loc_longitude&yField=loc_latitude&crs=EPSG:4326&spati
alIndex=no&subsetIndex=no&watchFile=no', 'INTERSECT' :
'D:\\r\\KenyaDSM\\data\\soil\\ke_sp\\soil_counties.shp', 'OUTPUT' :
'D:/r/KenyaDSM/data/soil/ke_sp/sample_sp.shp', 'PREDICATE' : [0,6] }


No spatial index exists for input layer, performance will be severely degraded

Execution completed in 0.50 seconds

Results:

{'OUTPUT': 'D:/r/KenyaDSM/data/soil/ke_sp/sample_sp.shp'}


Loading resulting layers

Algorithm 'Extract by location' finished
```

C.3. Clip Raster by Extent Function

```
QGIS version: 3.34.3-Prizren

QGIS code revision: 47373234ac

Qt version: 5.15.3

Python version: 3.9.18

GDAL version: 3.8.3

GEOS version: 3.12.1-CAPI-1.18.1

PROJ version: Rel. 9.3.1, December 1st, 2023

PDAL version: 2.6.0 (git-version: 3fced5)

Algorithm started at: 2024-03-18T22:26:14

Algorithm 'Extract by location' starting…

Input parameters:

{ 'INPUT' :
'delimitedtext://file:///D:/r/KenyaDSM/data/soil/ke_sp/ke_sp.csv?type=csv&maxFields
=10000&detectTypes=yes&xField=loc_longitude&yField=loc_latitude&crs=EPSG:4326&spati
alIndex=no&subsetIndex=no&watchFile=no', 'INTERSECT' :
'D:\\r\\KenyaDSM\\data\\soil\\ke_sp\\soil_counties.shp', 'OUTPUT' :
'D:/r/KenyaDSM/data/soil/ke_sp/sample_sp.shp', 'PREDICATE' : [0,6] }


No spatial index exists for input layer, performance will be severely degraded

Execution completed in 0.50 seconds

Results:

{'OUTPUT': 'D:/r/KenyaDSM/data/soil/ke_sp/sample_sp.shp'}


Loading resulting layers

Algorithm 'Extract by location' finished
```

C.4. Creating Data Partitions for Spatial Modelling

```
part <- createDataPartition(mg_dat$log_ExMg, times = 1, p = 0.2, list = FALSE)
mg_test <- mg_dat[part, ]
mg_train <- mg_dat[-part, ]
head(mg_train)
```

C.5. Conversion of Data to Spatial Format for Kriging

```
coordinates(k_train) <- ~ X + Y
k_train@proj4string <- CRS(wsg84)
k_train <- spTransform(k_train, CRS(utm))
k_covs <- projectRaster(k_covs, crs = CRS(utm))
k_covs <- as(k_covs, "SpatialGridDataFrame")
k_train <- k_train[which(!duplicated(k_train@coords)),] #remove any duplicates
```

C.6. Generation of Elbow Co-efficient Plot

```
clusters = list(range(2,16)) # test clusters from 2 to 15
ssd = [] # sum of squared distances

for k in clusters:
    km = KMeans(n_clusters=k, random_state=0, n_init="auto")
    km.fit(nutrients)
    ssd.append(km.inertia_)

# elbow plot
plt.figure(figsize=(15,7))
ax = sns.lineplot(x=clusters, y=ssd, marker="o", dashes=False)
ax.set_xlabel('Clusters')
ax.set_ylabel('SSD')
ax.set(xlim=(1,16))
plt.plot();
```

C.7. Creation of Map for Food Predictor

```
// create map with attributions
const map = L.map('map').setView([-0.5, 37.5], 9)

L.tileLayer('https://tile.openstreetmap.org/{z}/{x}/{y}.png', {
    maxZoom: 19,
    attribution: '&copy; <a
href="http://www.openstreetmap.org/copyright">OpenStreetMap</a>'
}).addTo(map);

const search = new GeoSearch.GeoSearchControl({
    provider: new GeoSearch.OpenStreetMapProvider(),
    style: 'bar',
  });
map.addControl(search); //add search control

// add polygon of soil region
const polygon = L.polygon([
    [-0.9147, 37.2602],
    [0.0688, 37.2602],
    [0.0688, 38.3019],
    [-0.9147, 38.3019]],{
        color: '#ff9505',
        fillColor: '#ffc971',
        fillOpacity: 0.1
    }).addTo(map);
```

VT OMNES VNVM SINT

C.8. Generation of Silhouette Co-efficient Plot

```
# silhouette score for verification
clusters = list(range(2,16)) # test clusters from 2 to 15
ssc = [] # sum of squared distances

for k in clusters:
    km = KMeans(n_clusters=k, random_state=0, n_init="auto")
    km.fit(nutrients)
    km_labels = km.labels_
    ssc.append(silhouette_score(nutrients, km_labels))

# elbow plot
plt.figure(figsize=(15,7))
ax = sns.lineplot(x=clusters, y=ssc, marker="o", dashes=False)
ax.set_xlabel('Clusters')
ax.set_ylabel('SSD')
ax.set(xlim=(1,16))
plt.plot();
```

VT OMNES VNVM SINT