Master Thesis

K.W.M. Coolen

October 4, 2024

Contents

1 Introduction

1.1 Population on Earth

In November 2022 the world reached a population of 8 billion people. The population is expected to increase by nearly 2 billion persons in the next 30 years, from 8 billion to 9.7 billion in 2050 and could peak at 10.4 billion in the mid-2080s. The growth has been driven by increasing numbers of people surviving to reproductive age, the gradual increase in human lifespan, increasing urbanization, and accelerating migration. Major changes in fertility rate have accompanied this growth. The average fertility was 2.3 births per woman in 2021, while it was 5 births per woman in 1950. The prediction is that the global fertility drops to 2.1 births per woman by 2050 (United Nations Department of Economic and Social Affairs, Population Division, 2022). It took 12 years to grow from 7 to 8 billion people, it is going to take around 15 years to reach 9 billion, which is a sign that the overall growth rate of the global population is slowing. Still, some countries have high levels of fertility. Countries with the highest fertility levels tend to be those with the lowest income per capita. As a result, global population growth has become concentrated among the world's poorest countries, most of which is in sub-Saharan Africa. More than half of population growth between now and 2050 is expected to occur in Africa, because it has the highest rate of population growth among major areas. The population of sub-Saharan Africa is projected to double by 2050 (United Nations, n.d.).

Figure 1: Diani Beach

Figure 2: Tsavo National Park

1.2 Food insecurity in Africa

Africa is currently facing a food crisis due to the effects of the war in Ukraine, climate change, economic issues, and the COVID-19 pandemic. Africa is not on track to meet the food security and nutrition targets of the Sustainable Development Goal 2 on Zero Hunger for 2030, and the Malabo targets of ending hunger and all forms of malnutrition by 2025. The Malabo Declaration on Accelerated Agricultural Growth and Transformation for Shared Prosperity and Improved Livelihoods is a set of new goals created by the African Union to achieve the agricultural vision for Africa to be achieved by 2025. Some of these goals include for instance allocating at least 10% of public resources to agriculture, ending hunger in Africa, reducing by poverty by 50%, and enhancing resilience of livelihoods and production systems (African Union, 2014). Recent estimates show that nearly 282 million people in Africa, which is about 20% of the population, were undernourished in 2022 and 868 million people were moderately or severely food insecure. The regions in Africa with the most undernourished people are Central and Eastern Africa (29.1% and 28.5%), which is respectively 57 million and 134.6 million. According to the FAO, AUC, ECA, and WFP, 2023, people face moderate food insecurity when they are uncertain of their ability to obtain food and have been forced to reduce, at times over the year, the quality and/or quantity of food they consume due to lack of money or other resources and severe food insecurity means that individuals have likely run out of food, experienced hunger and, at the most extreme, have gone for days without eating, putting their health and well-being at serious risk. The difference in numbers between food insecurity and undernourishment can be explained by their definitions. People experiencing moderate food insecurity might have issues with acquiring healthy nutritious food, but that does not mean that they are undernourished. This paper is taking a closer look at Kenya, a country in Eastern Africa. Kenya's prevalence of moderate or severe food insecurity had an average of 72.3% from 2020 to 2022, while the 2014-2016 average was 50.7%, so in 6 years, the prevalence increased quite a bit (FAO, AUC, ECA, and WFP, 2023). Thus, food security in Kenya has been decreasing the past decade and an increasing population will further deteriorate food security unless more actions are taken by the Kenyan government.

1.3 Geography of Kenya

Kenya has an area of 582,646 square kilometres and shares a border with Tanzania, Uganda, South Sudan, Ethiopia, Somalia, and the Indian Ocean. Kenya has diverse terrains. In the east, it is bordered by the Indian Ocean with white sand beaches, azure water, mangrove swamps, and palm trees as shown in Figure 1. In the north there is a small area of desert and semi-desert, which consists of sand-dunes, volcanic terrain, and forested mountains (Figure 4). On the western side of Kenya is the Rift Valley (Figure 3), which runs between Mozambique and Jordan and holds several lakes including Lake Turkana which is the world's largest permanent desert lake. Southwest of the Rift valley lies Lake Victoria, the second largest freshwater lake in the world. Furthermore, Kenya has dense rain forest and savanna grasslands as can be seen in Figure 2 (WorkingAbroad, n.d.). Figure 5 shows a map of Kenya containing geographical features such as elevation, rivers, lakes, major roads, mountain ranges, and other features.

1.4 Agriculture in Kenya

The agriculture sector is vital for Kenya, as it contributed around 21% of the Gross Domestic Product (GDP) and another 17.7% of GDP indirectly through linkages with other sectors in 2023 (KNBS, n.d.). In the agriculture sector works more than 40% of the population and more than 70% of the rural population. Agriculture accounts for 65% of the export earnings and provides livelihood for more than 80% of the Kenyan population, where livelihood means employment, income, and food security. Kenya has 28 million hectares of agricultural land which is around 48% of

Figure 3: Rift Valley

Figure 4: Suguta Valley

Figure 5: Map of Kenya showing major geographical features

Kenya's total land area, but only 21% of this area is arable land. A smallholder farmer has an average land size between 0.2 and 3 hectares, and together they provide 75% of Kenya's total agricultural output (Government of Kenya, n.d.). Kenya's main agricultural products are tea, coffee, maize, sugarcane, and horticulture. Some other products are wheat, sorghum, rice, beans, peas, green grams, and potatoes, livestock, fish, and dairy farming. The agricultural market is very important as a foreign-exchange earner, especially tea and cut-flowers. Fruits, vegetables, sisal, and cotton are also much exported products. Kenya is also the largest exporter of pyrethrum, a flower that is used to create the pesticide pyrethrin. Coffee was an important export product but has been declining since the 90s. Maize, wheat, sugarcane, dairy goods, and livestock are primarily for the domestic markets. Agriculture is largely subsistence (grow crops for own use) and productivity has stagnated in recent years, despite population growth. Recent crises such as drought, war in Ukraine, and Covid have deteriorated the vulnerability of basic livelihoods, which has a negative impact on food security. Most of the land in Kenya is semi-arid or arid and is used for pastoralism and the remaining land which is suitable for farming do not achieve maximum yields (USAID). This is because many farmers work without updated technology or basic agricultural inputs and they do not have adequate financial or extension services. Other reasons for lower yields are the occurrence of crop diseases and pests, bad access to storage and preservation facilities, the lack of infrastructure in rural areas of Kenya, and the dependence on rain-fed farming systems (FAO, World Bank, Alliance of Bioversity International and CIAT, n.d.).

1.5 Agro-Ecological Zones

"Agro-ecological Zoning (AEZ) refers to the division of an area of land into smaller units, which have similar characteristics related to land suitability, potential production and environmental impact. An Agro-ecological Zone (AEZ) is a land resource mapping unit, defined in terms of climate, landform and soils, and/or land cover" (FAO, 1996). An AEZ has a specific range of potentials and constraints for land use which is defined by the crop-growing period, temperature regime and soil type.

Kenya's agro-ecological zoning is mainly based on altitude, rainfall and potential productivity of crops and livestock. Kenya is divided into low, medium, and high potential areas for agricultural productivity, which gives a general indication how land can be used in specific areas. Kenya has 7 distinct AEZs, where AEZs I and II are high potential areas, AEZs III and IV correspond to medium potential areas, and AEZs V, VI, and VII represent low potential areas (picture to make it clearer, use screenshot from arcgis). The low potential AEZs have arid and semi-arid conditions and occupy about 84% of the country's total land mass. Each AEZ faces land degradation, which is orchestrated by deforestation and erosion of soil by wind and water (Wanjira et al., 2020).

1.6 Kenyan soils

Kenya has a wide range of soils due to variation in geology, in relief and climate. Soil resources vary from sandy to clayey, shallow to very deep and low to high fertility. However, most soils have serious limitations such as salinity, sodicity, acidity, fertility and drainage problems. Some major soils in Kenya are ferralsols, vertisols, acrisols, lixisols, luvisols, nitisols, andosols, alisols, and planosols. Andosols, young volcanic soils, are porous, have a high water-storage capacity and a low bulk density and occur in areas with steep slopes and high rainfall. They are also acidic due to high levels of aluminium and the high leaching of soluble bases. Besides, andosols are susceptible to erosion as they mostly occur on steep slopes. The soils are mainly used for tea, pyrethrum, temperate crops, and dairy farming. Nitisols occur in highlands and on volcanic slopes and are developed from volcanic rocks and have better physical and chemical properties than other tropical soils. Most nitisols are acidic because of the leaching of soluble bases. Nitisols are the best agricultural soils found in the region and are used for plantation crops and food production, such as banana, tea, and coffee. Acrisols, alisols, lixisols, luvisols are soils that occur in the coffee zones in the sub-humid areas. They have a relatively low water-storage capacity, compared with nitisols. Acrisols and alisols have a low pH in wet areas, aluminium and manganese toxicities and low levels of nutrients and nutrient reserves. The 4 soils have poor structure and need erosion control measures. Ferralsols are old, highly weathered and leached soils, and the topsoil has poor fertility, whereas the subsoil has low cation exchange capacity. Phosphorus and nitrogen are always deficient, but ferralsols have a lot of aluminium and iron. Ferralsols are used to grow annual and perennial crops and are suitable for oil palm, rubber and coffee. Planosols and vertisols occur mostly in rice growing areas and are found in semi-arid and sub-humid environments. Due to the high clay content in the subsoil, this layer is impermeable in the subsoil resulting in a slow vertical and horizontal poor drainage and poor workability of the soils (Infonet Biovision, n.d.).

1.7 Fertilizer in Kenya

The last few years, soil health in Kenya has been declining and according to Kenya's agriculture ministry, 63% of arable land is acidic. But there are farmers who double their fertilizer application if they experience low yields during a season. If the soil pH is acidic, then plants cannot absorb the existing minerals in the soil and adding more fertilizer

exacerbates the problem, which leads to more soil degradation. The most common option to raise soil pH is to lime the soil, which means that calcium and magnesium rich materials are applied to the soil. The soil degradation is due to the accumulation of fertilizer metals in the soil because specific fertilizers are overused which changes the soil pH to acidic. That is why soil testing is important to determine pH levels before deciding which kind of fertilizer to use. There is not enough soil data, so the wrong fertilizers are used by farmers in an attempt to increase yield and it has an opposite effect from what is required (AfricaNews, 2024). However, a major challenge is optimizing fertilizer use across its diverse agro-ecological zones. The Kenyan government has implemented several fertilizer subsidy programs over the years to boost food production, which were not as successful as hoped. The fertilizer subsidy programs in Kenya are not effective because of various reasons, such as

- Kenyan smallholder's adoption of modern agricultural technologies has remained persistently low for decades (Bueno, 2024).
- Farmers that are familiar with fertilizer do not use the subsidized offering (Bueno, 2024).
- A lack of knowledge by farmers, which leads to them using the wrong fertilizers, because most do not know which micro nutrients are lacking in the soils (Omondi, 2023).
- Fertilizer has been used in a non-optimal way. An example of this is that farmers were asked to switch from Di-ammonium Phosphate (DAP) to NPK fertilizer due to increased acidity of the soil, but most farmers continued using one bag of NPK per acre as they used to do with DAP instead of two (Omondi, 2023).
- Other reasons are limited access, delayed fertilizer delivery to depots/stores, long queues when picking up the fertilizer, lack of communication, only one type of fertilizer available, high transportation costs due to a low number of fertilizer distribution points and increasing fuel prices (Bueno, 2024), etc.

1.8 Objective

The objective of this research paper is to enhance fertilizer optimization in Kenya through a data-driven approach, utilizing the latest advancements in Data Science and Artificial Intelligence. The primary focus is on developing a comprehensive strategy to assist the Kenyan government in implementing effective fertilizer subsidy programs. By integrating diverse soil health data across various agro-ecological zones, the project seeks to address the unique nutrient requirements of different regions, ultimately leading to improved agricultural productivity and sustainability.

In order to fulfill the objective, we start with developing a soil sampling plan that uses both wet chemistry methods and satellite imagery-based techniques. This hybrid approach aims to capture a detailed and accurate picture of soil health which highlights nutrient gaps across different agro-ecological zones in Kenya in the form of a soil health map. This map is going to be created using Digital Soil Mapping (DSM) with ML techniques such as Random Forest or Artificial Neural Network. The soil health map will serve as a foundational tool for understanding regional soil health status. Using the soil health map, optimized fertilizer mixes tailored are recommended for different sub-counties/regions based on specific crop types prevalent in each area. The dataset will be analysed using Machine Learning/Artificial Intelligence, which enables the generation of data-driven recommendations. The recommendations are aimed at informing policy-level decisions, to help the government distribute fertilizer subsidies more effectively and efficiently. The expectation is that this approach enhances crop yields and soil health. Furthermore, the approach supports sustainable farming practices, which aligns with Kenya's agricultural and environmental goals.

The rest of the paper is structured as follows: chapter 2 talks about the methods and data used in the study, chapter 3 describes the results, such as the optimal sample size, the sample locations, and the soil health map, chapter 4 contains the discussion and further research, chapter 5 is the conclusion of the project. (just the beginning, needs a bit more detail, look at other papers for examples)

2 Literature Review

2.1 Sample size

Two important parts of a soil sampling plan are the sample size and the sampling locations. First, the sample size is calculated which determines the number of sampling locations/plots. The sample size usually depends on the size of the study area as in the larger the area, the more samples are required. Calculating the optimal size is useful for a soil sampling operation, because collecting too many samples can lead to oversampling which results in having observations that do not provide additional information for the target soil properties. Besides, soil sampling and subsequent soil research in a laboratory is expensive, so collecting too many samples is wasteful. (Saurette et al., 2022) state that for many DSM projects, the financial considerations override what is deemed as the optimal sample size. That is why they think it should be considered to determine a minimum sample size that would be required for a DSM project. Their

paper proposes a technique to determine the minimum sample size for the cLHS algorithm based on the interpretation of histograms of the environmental covariates. Another method to calculate the sample size is demonstrated in (Malone et al., 2019). They determined an optimal sample size by comparing the empirical distribution of the samples and the population with increasing sample size using Kullback-Leibler divergence (KLD). This results in a exponential growth curve that plateaus at a point and then a rule of diminishing returns is used to determine an optimal sample size. (Stumpf et al., 2016) determined the optimal sample size by comparing the variance of the covariate space of the study area to the variance of the original sample sets with multiple sizes. (Wu et al., 2022) used random forest to predict SOC content for 10 different calibration sample sizes and identified the optimal sample size by checking which predictive model had the lowest root mean squared error and the largest explained variance.

2.2 Sampling locations

Both the sample size and the sample locations form an integral part of the soil sampling plan. After establishing the sample size, the sample locations can be determined by selecting a sampling design. Brus, 2019 noted that there is no single best sampling design, and that the best design depends on the technique used for mapping. The sampling design type determines how sampling locations are selected. Possible design types are simple random sampling (SRS), stratified random sampling, two-stage cluster random sampling, spatial coverage sampling (SCS), conditioned Latin hypercube sampling (cLHS), regular grid sampling, feature space coverage sampling (FSCS), etc where each design type has its advantages and disadvantages. (Saurette et al., 2024) determined sample locations using a combination of expert opinion sampling, selected sample points using the cLHS algorithm, and opportunistic sampling when cLHS locations were inaccessible, whereas (Tziachris et al., 2019) used simple random sampling for selecting sample locations. There is also the possibility of using legacy soil data points, which means no new soil sample locations have to be determined. (Wadoux et al., 2019) used the sampling points from the European Land Use/Cover Area frame Statistical Survey (LUCAS) to investigate what makes a sampling design optimal for mapping using random forest (RF). They estimate the population mean squared error (MSE) with different sampling designs like SRS without replacement, SCS, cLHS, FSCS, and MSE optimized. Table 1 in (Wadoux et al., 2020) shows 75 recent case studies of DSM which were produced using a ML algorithm. The table contains amongst others the sampling design used in the study. Most studies do not specify the sampling design of their study and it is the theory is that in these cases the sample comes from several sources, like expert-based designs, legacy data, and combination of surveys and the sources all had a different sampling design. From the papers that did specify, non-probability sampling designs such as cLHS and grid-based are the most used for (model) calibration. Probability sampling design are used in 25 % of the studies.

2.3 Data Soil Mapping (using ML)

(Lagacherie and Mcbratney, 2006) defined DSM as "the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observations and knowledge and from related environmental variables." (Hence), The most common research/work in DSM is aimed at predicting soil class or soil attribute from the scorpan factors using a spatial soil prediction function with autocorrelated error. (McBratney et al., 2003) reformulated Jenny's clorpt model into the scorpan model which has the following equation for mapping soil:

$$S = f(s, c, o, r, p, a, n) + e \tag{1}$$

The scorpan model has the same variables as the clorpt model and some additional ones. S represents a soil attribute or a class to be predicted using a quantitative function f which uses all the scorpan variables are a combination of them. The s stands for (previously) measured intrinsic properties of the soil, c is climate, o are organisms, r is relief/topography, p is parent material, a is time/age, n represents the spatial location/coordinates of a sample, and e is spatially correlated residuals.

DSM has a dual objective. The first objective is gaining better understanding of mechanistic soil process and the second objective is making accurate predictions. The current literature on ML for DSM has focused almost solely on the second objective. In many situations it is important that the soil map is accurate, for example in policy making. That is why there has been an increase in the number of publications where the main interest was prediction of a soil property or class (Wadoux et al., 2020). The increase of DSM related activities is caused by the convergence of several factors. Firstly, there is a worldwide demand for qualitative good digital soil maps and for quantitative and spatial soil information. Also, there is more digital spatial data available and there are more databases containing measured/inferred soil properties combined with known environmental variables. Furthermore, the computing power for processing large datasets has increased and became more available and there was development of algorithms, data mining tools and GIS. The limited financial and labor resources for producing soil maps and a new generation of soil scientists are also beneficial (Wadoux et al., 2020, Minasny and McBratney, 2016, Khaledian and Miller, 2020). There is a need for up-to-date and fine resolution soil data, because it is needed to address natural resource management problems like climate change, food and water security, and biodiversity management. Many of the developments in

DSM originated from universities, because they were unable to create soil maps in the past as it was the domain of national soil survey centres.

The DSM community has shifted its attention from geostatistical methods as the go-to method to using ML algorithms as main mapping tool for spatial prediction. Geostatistical models are often used in soil mapping because they have several advantages, such as that the model assumes a statistically sound model for spatial variation, spatial autocorrelation is explicitly modelled, and an explicit measure of the uncertainty is associated with the prediction. On the other hand, geostatistical mapping has some disadvantages as well. For one, the residuals are assumed to be normally distributed, stationary, and isotropic. Also, it has a hard time modelling the non-linear relationship between a soil property and cross-correlated covariates and in heterogeneous areas, the model of spatial variation fails to capture gradual and sudden changes in soil variation. Lastly, a geostatistical model with a large sample size and/or number of prediction locations is computationally demanding (Wadoux et al., 2020). In the 1990s, machine learning (ML) emerged as a new tool for DSM and as an alternative for geostatistical models. ML algorithms do not make an assumption of the observations' distribution, unlike geostatistical methods where transformation of the original observations is often required to satisfy the assumption. Geostatistical models and ML algorithms have a different purpose when applied in DSM. ML algorithms mostly emphasize prediction accuracy whereas geostatistical models attempt to infer the process which generated the data through a pre-defined model of spatial variation (Wadoux et al., 2020).

The number of research papers on DSM has grown the last two decades and the proportion of these papers using ML has increased over the years. Other used methods in DSM are kriging and hybrid. Fig 2 in (Khaledian and Miller, 2020) shows the number of research papers and books on DSM using popular ML algorithms, such as KNN, RF, Cubist, ANN, multiple linear regression (MLR), and support vector regression (SVR). It can be seen that in 2005 only ANN was used, but in 2018 all 6 methods were applied. RF has the highest growth rate of the methods while Cubist, SVR, and KNN started being used around 2013. RF and Cubist are decision trees models and these models can handle linear and non-linear relationships, which is why they have been getting more attention. Furthermore, tree models outperform classical models like MLR and they require little pre-processing. Other ML algorithms which have been used in DSM research are classification and regression trees (CART), bagging, quantile RF, boosted regression tree, convolutional neural network (CNN), support vector machines (SVM), k-nearest neighbours (KNN), and generalized boosted regression. Table 1 in (Wadoux et al., 2020) shows a summary of 75 case studies of DSM which were produced using a ML algorithm. The table contains amongst others the sampling design, the sample size, and the applied ML model(s). Most of the ML algorithms methods mentioned previously are present in the table, which shows that there are many possibilities for creating a soil health map. For example, (Blanco et al., 2018; Shi et al., 2018; Wiesmeier et al., 2011) used RF for mapping.

Thus, ML algorithms are widely used in DSM for regression and classification, where a high percentage is aimed at prediction. The accuracy from these models are higher, because ML algorithms are not constrained by a pre-defined model of the soil spatial variation, whereas geostatistical and mechanistic models are. When a DSM project aims to predict the spatial distribution of soil target properties, an approach has to be chosen. Every algorithm has its strengths and weaknesses, so there is not a ML algorithm that performs best for all problems. Some algorithms are more suitable than others depending on the nature of the data and purpose of mapping activity. Selecting an appropriate algorithm for a problem depends on several factors such as the algorithm's ability to select covariates, the quantity of available soil samples, the efficiency of the computational process, the algorithm's hyperparameters, and interpretability of the created model (Khaledian and Miller, 2020).

3 Methodology

3.1 Data

3.2 Optimal Sample Size

For every region/sub-county in Kenya, we want to determine the minimum number of soil samples to capture at least 90-95% of variability within the environmental covariates. Divergence metrics originate from information theory and are used to compare two probability distribution functions. Several divergence metrics are Kullback-Leibler divergence, Jensen-Shannon divergence, squared Euclidean distance, and Bhattacharyya distance. The Jensen-Shannon divergence (DJS) is a better divergence metric then Kullback-Leibler divergence (DKL) **(source)**. We then draw samples from the environmental covariate rasters and compare to the distribution of each covariate in the sample plan to that of the full covariate (population) of the study area using the Jensen-Shannon divergence across an increasing sample size (Saurette et al., 2024). The sampling plans can be created using sampling algorithms, such as conditioned Latin Hypercube Sampling (cLHS), Feature Space Coverage Sampling (FSCS), Spatial Coverage Sampling (SCS), Simple Random Sampling (SRS), etc. The idea is to create 5 to 10 independent sample plans with increasing sample sizes, where the actual sample sizes depend on the number of available data points. Finally, a cumulative distribution

function of the DJS is used to determine the optimal sample size (Saurette et al., 2024). There does not seem to be much difference in the results between the sampling algorithms, so anyone of them could be used.

3.3 Soil Sampling Plan

Possible aims of the soil sampling plan are sampling for estimating means and totals, sampling for mapping, and sampling for monitoring.

First we decide what sampling method is needed for the soil sampling plan. There are three options, namely convenience sampling, purposive sampling, and probability sampling. With convenience sampling, samples are selected based on convenience, for example areas near road sides for easy access. Purposive sampling selects samples such that a given purpose/objective is best served. Locations can be selected based on their characteristics derived from the environmental variables. With probability sampling, the selection of sampling locations is done using a sampling design to ensure that the inclusion probabilities are known for each sampling location. (more suited for methodology) After picking a sampling approach/method, a design type is chosen. A design type defines how sampling locations are selected.

3.4 Soil health map

DSM requires environmental layers/variables, geospatially referenced soil data and a model that characterizes the relationship between the soil data and the environment to make predictions for unsampled locations (Heung et al, 2021). Digital Soil Mapping (DSM) can be defined as the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge and from related environmental variables (Lagacherie et al., 2006). Developing a soil map using DSM techniques largely requires the following three components according to Heung et al, 2021:

1. a suite of environmental layers (i.e., predictors or covariates) that represent the scorpan factors.
2. soil data that are geospatially referenced.
3. a model that characterizes the relationship between the soil data and the environment to make predictions for unsampled locations.

Environmental covariates are derived from digital elevation maps, satellite data, images, landcover, climatic/geological data, soil maps, expert knowledge from scientists/technicians. Examples of environmental covariates are slope, terrain, elevation, solar radiation, slope height, slope length, topographic wetness index, roughness, annual average enhanced vegetation index, average temperature, average rainfall. The environmental covariates are known to influence the spatial distribution of target soil properties or classes (Saurette et al., 2024). The chosen prediction model is fit using a target soil property as dependent variable and the covariates as predictors. The soil property is mapped using the prediction model and assessed for quality and uncertainty. Target soil properties for data soil mapping can be pH, soil organic carbon, macro- and micro nutrients (N, P, K, Zn, B, Mn, Mg, Ca, Fe, Cu, Cl, S, Mo), organic matter, cation exchange capacity, texture, bulk density. In this study the focus is on creating soil health maps for N, P, K, and pH, because they are the main ingredients for fertilizer or heavily influenced by fertilizer use. For mapping using Machine Learning (ML) techniques with covariates, (Brus, 2019) recommends selecting the sample using FSCS or cLHS. Both cLHS and FSCS aim for an even sampling density in the multivariate feature space, but in different ways. In cLHS it is done through minimization of a criterion which is a function of the marginal distributions and correlation matrix of the covariates using spatial simulated annealing, in FSCS it is achieved through minimization of a feature space distance criterion between sampling and prediction points using the k-means algorithm (Wadoux et al., 2019). There are several valid options for sampling designs and mapping techniques. To decide which one to use, a part of the data is used to implement multiple methods to compare which design/technique performs the best. The data can be split into a training, validation, and test set. For example, the division could be 60% training, 20% validation, 20% test. The training set is used to train the models and cross-validation is used to find the best model/hyperparameters. The test set is then used on the best model. Methods that can be used for mapping are random forest, support vector machine, ANN/deep learning, multinomial regression, RFspatial, gradient boosting, cubist, and k-nearest neighbours. Cross-validation can be used to avoid overfitting. The size of a validation set depends on the amount of hyperparameters a method has. If there is not enough data, this does not work. This process should result in an accurate soil health map for one or more soil target properties with sampling locations and an optimal sample size. The soil health map is used as a tool for understanding regional soil health status.

4 Results

4.1 Optimal Sample Size

4.2 Soil Sampling Plan

4.3 Soil health map

5 Discussion

6 Conclusions

References

African Union. (2014). *Malabo declaration on accelerated agricultural growth and transformation for shared prosperity and improved livelihoods*. Retrieved September 11, 2024, from https://www.resakss.org/sites/default/files/Malabo%20Declaration%20on%20Agriculture_2014_11%2026-.pdf

Blanco, C. M. G., Gomez, V. M. B., Crespo, P., & Ließ, M. (2018). Spatial prediction of soil water retention in a páramo landscape: Methodological insight into machine learning using random forest. *Geoderma*, *316*, 100–114.

Brus, D. (2019). Sampling for digital soil mapping: A tutorial supported by r scripts. *Geoderma*, *338*, 464–480.

FAO. (1996). *Agro-ecological zoning guidelines*. Retrieved September 17, 2024, from https://www.faoswalim.org/resources/Agriculture/AGRO-ECOLOGICAL_ZONING_Guidelines.pdf

FAO, AUC, ECA, and WFP. (2023). *Africa - regional overview of food security and nutrition 2023:statistics and trends*. Retrieved September 11, 2024, from https://doi.org/10.4060/cc8743en

FAO, World Bank, Alliance of Bioversity International and CIAT. (n.d.). *Digitial agriculture profile - kenya*. Retrieved September 13, 2024, from https://openknowledge.fao.org/server/api/core/bitstreams/ad4161f5-9fdf-4b9c-bcf6-8f473f929af1/content

Government of Kenya. (n.d.). *Agricultural sector development strategy 2010-2020*. Retrieved September 13, 2024, from https://faolex.fao.org/docs/pdf/ken140935.pdf

Heung et al. (2021, November). *3.4: Digital soil mapping*. Retrieved September 11, 2024, from https://geo.libretexts.org/Bookshelves/Soil_Science/Digging_into_Canadian_Soils%3A_An_Introduction_to_Soil_Science/03%3A_Digging_Deeper/3.04%3A_Digital_Soil_Mapping

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, *81*, 401–418.

KNBS. (n.d.). *2023 economic survey*. Retrieved September 13, 2024, from https://www.knbs.or.ke/wp-content/uploads/2023/09/2023-Economic-Survey.pdf

Lagacherie, P., Mcbratney, A., Voltz, M., Grunwald, S., Ramasundaram, V., Comerford, N., & Bliss, C. (2006). Digital soil mapping. In *Digital soil mapping* (p. iii, Vol. 31). Elsevier. https://doi.org/https://doi.org/10.1016/S0166-2481(06)31060-4

Lagacherie, P., & Mcbratney, A. B. (2006). Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. *Developments in soil science*, *31*, 3–22.

Malone, B. P., Minansy, B., & Brungard, C. (2019). Some methods to improve the utility of conditioned latin hypercube sampling. *PeerJ*, *7*, e6451. https://doi.org/https://doi.org/10.7717%2Fpeerj.6451

McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1-2), 3–52.

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

Saurette, D. D., Biswas, A., Gillespie, A. W., et al. (2022). Determining minimum sample size for the conditioned latin hypercube sampling algorithm. *Pedosphere*. https://doi.org/https://doi.org/10.1016/j.pedsph.2022.09.001

Saurette, D. D., Heck, R. J., Gillespie, A. W., Berg, A. A., & Biswas, A. (2024). Sample size optimization for digital soil mapping: An empirical example. *Land*, *13*. https://doi.org/10.3390/land13030365

Shi, J., Yang, L., Zhu, A.-X., Qin, C., Liang, P., Zeng, C., & Pei, T. (2018). Machine-learning variables at different scales vs. knowledge-based variables for mapping multiple soil properties. *Soil Science Society of America Journal*, *82*(3), 645–656.

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., Wadoux, A., Xiang, W., & Scholten, T. (2016). Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. *Journal of Plant Nutrition and Soil Science*, *179*(4), 499–509. https://doi.org/https://doi.org/10.1002/jpln.201500313

Tziachris, P., Aschonitis, V., Chatzistathis, T., & Papadopoulou, M. (2019). Assessment of spatial hybrid methods for predicting soil organic matter using dem derivatives and soil parameters. *Catena*, *174*, 206–216. https://doi.org/https://doi.org/10.1016/j.catena.2018.11.010

United Nations. (n.d.). *Population*. Retrieved September 11, 2024, from https://www.un.org/en/global-issues/population

United Nations Department of Economic and Social Affairs, Population Division. (2022). *World population prospects 2022: Summary of results*. Retrieved September 11, 2024, from https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf

Wadoux, A. M. J.-C., Brus, D. J., & Heuvelink, G. B. M. (2019). Sampling design optimization for soil mapping with random forest. *Geoderma*, *355*. https://doi.org/https://doi.org/10.1016/j.geoderma.2019.113913

Wadoux, A. M.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, *210*, 103359.

Wanjira, E., Muriuki, J., & Ojuok, I. (2020). *Farmer managed natural regeneration in kenya*. Retrieved September 17, 2024, from https://fmnrhub.com.au/wp-content/uploads/2022/05/FMNR-KENYA-MANUAL.pdf

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

WorkingAbroad. (n.d.). *Kenya - geography and climate*. Retrieved September 11, 2024, from https://www.workingabroad.com/travel/kenya-geography-and-climate/

Wu, T., Wu, Q., Zhuang, Q., Li, Y., Yao, Y., Zhang, L., & Xing, S. (2022). Optimal sample size for soc content prediction for mapping using the random forest in cropland in northern jiangsu, china. *Eurasian Soil Science*, *55*(12), 1689–1699. https://doi.org/https://doi.org/10.1134/S1064229322600816