Plan of approach for creating soil health maps

Kelt Coolen (u887571)

Tilburg University

What is the problem?

Kenya's agricultural sector is vital for its economy and food security. The last few years, the soil health in Kenya has been declining and according to Kenya's agriculture ministry, 63% of arable land is acidic. The most common option to raise soil pH is to use liming. 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. That is why soil testing is important to determine pH levels before deciding which kind of fertilizer to use. However, a major challenge is optimizing fertilizer use across its diverse agro-ecological zones. Traditional approaches may not efficiently address the varying soil health needs across regions. This project proposes leveraging Data Science and AI to create a large-scale soil health map, identifying nutrient deficiencies of the soil and optimizing fertilizer mixes at a sub-county level to improve agricultural productivity and sustainability. This leads to better knowledge for farmers regarding their soil properties and fertilizer subsidies can be targeted for regions/sub-counties. The primary focus is on developing a comprehensive strategy to assist the Kenyan government in implementing effective fertilizer subsidy programs.

The existing 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 micronutrients 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.

We want to assist the Kenyan fertilizer program by creating an accurate soil health map, but recent soil data is needed for better accuracy if available legacy data is not reliable. Because soil sampling is expensive, the number of soil samples should be minimized. Solving this sub-problem leads to an accurate soil health map, which will be used to propose fertilizer recommendations for different regions. These recommendations could help with points 3, 4, and 5 above to get a more effective fertilizer subsidy program.

Data

The data needed to perform is obtained from environmental covariates. One of the covariates can be derived from remote sensing data. There are several publicly available remote sensing-based covariates, such as:

1. SRTM and/or ALOS W3D Digital Elevation Model (DEM) at 30 m and MERIT DEM at 100 m (these can be used to derive at least 8–12 DEM derivatives of which some generally prove to be beneficial for mapping of soil chemical and hydrological properties).
2. Landsat 7, 8 satellite images, either available from USGS's GloVis / EarthExplorer, or from the GlobalForestChange project repository (Hansen et al. 2013).
3. Landsat-based Global Surface Water (GSW) dynamics images at 30 m resolution for the period 1985–2016 (Pekel et al. 2016).
4. Global Land Cover (GLC) maps based on the GLC30 project at 30 m resolution for 2000 and 2010 (Chen et al. 2015) and similar land cover projects (Herold et al. 2016).
5. USGS's global bare surface images at 30 m resolution.
6. JAXA's ALOS (PALSAR/PALSAR-2) radar images at 20 m resolution (Shimada et al. 2014); radar images, bands HH: -27.7 (5.3) dB and HV: -35.8 (3.0) dB, from the JAXA's ALOS project are especially interesting for mapping rock outcrops and exposed bedrock but are also used to distinguish between bare soil and dense vegetation. (Hengl, T. & MacMillan, R, 2019)

Sample Size

For every region/sub-county in Kenya, we want to determine the minimum number of soil samples to capture at least 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). 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.

Sample Locations

After calculating the sample size, the sample locations are determined. Again, a sampling design needs to be chosen. Some sampling designs are SCS, SRS without replacement, cLHS, FSCS, Artificial Neural Network (ANN) using Simulated Annealing (SA), Stratified Random Sampling. Each method has its advantages and disadvantages, but there is no single best sampling design, and the best design depends on the technique used for mapping.

It is possible that some sample locations cannot be accessed or sampled. That is why replacement points or replacement areas are created to act as substitutes for these inaccessible points. In some areas access is limited or restricted, for example conflict regions, steep slopes, remote areas, and natural protected areas. It is hard and possibly expensive to take a soil sample at these areas, so they should be excluded from the sampling area. Besides, in some of these places the soil cannot be cultivated, so sampling here would be irrelevant.

Mapping

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 and McBratney, 2007).

Developing a soil map using DSM techniques largely requires the following three ingredients:

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. (Heung et al, 2021)

Scorpan stands for spectral, climate, organisms, relief, parent material, time/age, location. 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 micronutrients (N, P, K, Zn, B, Mn, Mg, Ca, Fe, Cu, Cl, S, Mo), organic matter, cation exchange capacity, texture, bulk density.

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. Building on this, the project will

propose optimized fertilizer mixes tailored for the Kenyan sub-counties. These recommendations will be based on specific crop types prevalent in each area and the corresponding soil health data and other environmental variables. The recommendations are aimed at informing policy-level decisions, helping the government to distribute fertilizer subsidies more effectively and efficiently. This approach not only promises to enhance crop yields and soil health but also supports sustainable farming practices, aligning with Kenya's agricultural and environmental goals. The soil health maps could also be used for other areas, such as land management and environmental planning.

The research question is “how can fertilizer government subsidy programs in Kenya be made more efficient and effective using recommendations generated by a data-driven (ML/AI) approach?”

Literature

Alexandre M.J-C. Wadoux, Dick J. Brus, Gerard B.M. Heuvelink, Sampling design optimization for soil mapping with random forest, Geoderma, Volume 355, 2019, 113913, ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2019.113913.

Saurette, D.D.; Heck, R.J.; Gillespie, A.W.; Berg, A.A.; Biswas, A. Sample Size Optimization for Digital Soil Mapping: An Empirical Example. Land 2024, 13, 365. https://doi.org/10.3390/land13030365

Shao, Shuangshuang & Su, Baowei & Zhang, Yalu & Gao, Chao & Zhang, Ming & Zhang, Huan & Yang, Lin. (2022). Sample design optimization for soil mapping using improved artificial neural networks and simulated annealing. Geoderma. 413.115749. https://doi.org/10.1016/j.geoderma.2022.115749

Omondi, D. (2023, January 10). *Data cast doubt on success of Kenya’s fertiliser subsidy*. Business Daily. https://www.businessdailyafrica.com/bd/data-hub/data-cast-doubt-on-success-of-kenya-s-fertiliser-subsidy--4082078#google_vignette

Bueno, T. (2024, February 3). *Kenya’s Fertilizer Subsidy program supports farmers amidst economic challenges*. Fertilizer Daily. https://www.fertilizerdaily.com/20231229-kenyas-fertilizer-subsidy-program-supports-farmers-amidst-economic-challenges/

Heung, B., Saurette, D., & Bulmer, C. (2021, November 25). 3.4: Digital Soil Mapping. Geosciences LibreTexts. 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

Hengl, T., MacMillan, R.A., (2019). **Predictive Soil Mapping with R**. OpenGeoHub foundation, Wageningen, the Netherlands, 370 pages, www.soilmapper.org, ISBN: 978-0-359-30635-0.