Supporting the adoption of sustainable farming practices in western Kenya through data analysis.

Hack4Good Project, Fall 2021

Hack4Good
Apply your data science skills to create real-world impact

Oscar Pitcho, Nando Metzger, Afshan Anam Saeed and **Antoine Basseto**
ETH Zurich - Swiss Federal Institute of Technology, in collaboration with GIZ.
{opitcho, metzgern, afsaeed, abasseto}@ethz.ch

Abstract

The GIZ is a private German development agency on a mission to create a future worth living around the world. It partners with local stake-holders, public and private, and provides its know-how, expertise, and network to support development goals in all its programs. In western Kenya the GIZ leads the ProSoil program to increase farmer yield and preserve soil health. This enables affected farmers to have a long-term growth on their yield and sequestrate the carbon within the soil, reducing the regions CO2 emissions. The GIZ drives these goals by recommending and supporting the adoption of sustainable farming practices (e.g., use of organic soil fertilizer or crop rotations) by small farm-holders in the region.

1 Introduction

For reporting purposes, the organization conducts surveys on farms affected by the program. We were given these surveys, which fall in three categories: Adoption Surveys recording the adoption of farming practices by farming households, Farm-level Yield surveys, and soil studies of the region. These were conducted manually in western Kenya, consequently the data contained a significant amount of noise we had to correct. We assisted the organization by analyzing the data and attempted to build multiple models: a recommendation engine targeted for small farmers, a predictor of the effect of farming practices on yield, and clustering of the farming practices. We present all these solutions and their outcomes in this report. Alongside, we present all our data findings, our technical approach, and recommendations for future data collection.

2 Objectives and Goals

We identified three core challenges faced by the GIZ, and we present them below alongside our proposed solution:

1. **Scalability**: Experts advising individual farmers must interrogate the farmers, collect soil samples, and analyze these in the lab. The process is costly and lengthy and cannot scale to individual farms.

   Proposed Solution: We attempted to build a recommendation engine emulating the experts' reasoning: automating recommendations would accelerate the process and reduce costs.

2. **Adoption**: Recommended farming practices adoption should be improved. Predicting the effect sustainable farming practices would have on final yield would enable the GIZ to communicate tangible benefits to local farmers.

   Proposed Solution: We attempted to build a model predicting farming practices' impact on the yield using regression models.

3. **Effectiveness**: Although the GIZ recommends farming practices in groups, surveys show they are not applied following the same groups - farmers seem to pick and choose what they apply.

   Proposed Solution: We created a clustering model between different farming practices to spot patterns and understand how farmers combine practices in the field.

3 Implementation

To implement our solutions, we decided to use the framework Kedro (Bălan et al., 2020). It allows for the definition of pipelines with clean and modular code production.

We divided our pipelines into multiple steps: data cleaning, data-merging, data-enrichment, and data-science where model training and evaluation occur. For a complete view of the pipeline run `kedro viz`, more details are available in the readme of the git project.

3.1 Data-Cleaning

The datasets underwent extensive cleaning. Many features contained values with identical meaning but different values (for example, "pure_cropped" and "pure__cropped").

We dropped certain poor quality but essential features (e.g., variety of crops and the type of fertilizer used) we could not clean to a usable state. Moreover, they contained many categories compared to the other features and the total number of rows. Using them in a model would have led to a dramatic increase in dimensionality - we discarded them; we make a recommendation regarding these categorical features in section 5. Additionally, yield units do not have proper unit information. For example, the total yield can be in total weight or bags: the difference is not specified in the features (Adoption vs. Yield, for example).

Finally, up to 25% of rows did not contain farm coordinates. To overcome this issue and not discard these rows, we relied on a dataset of all Kenyan sub-counties and their geographical boundaries. By matching the sub-county of a row to a sub-county in our new dataset, we were able to assign the geographical center of the sub-county to datapoints that had no geographical coordinates provided.

3.2 Data-Merging

Averaging 400 farms per survey, it was impossible to use them separately to build a model, there were insufficient rows. So, we merged the datasets at our disposition to create larger datasets on which we could run our models. The goal was to sacrifice certain features in exchange for more data and thus be able to build robust models.

However, during merging, we had to drop many vital attributes such as household information, modes of learning of farming practices, reasons for disregarding a practice - we discarded the features not shared across datasets keeping only the ones present in all.

3.3 Data-Enrichment

As an intermediate step for further analysis, we designed data enrichment steps to add valuable features to our available data points. We grouped these in two categories:

3.3.1 Climate and Topological Data

The data provided did not contain any topological or climate information. We overcame this by designing a framework to enrich geospatial data with publicly available datasets. More specifically, we designed a system to enrich datasets containing WGS84 coordinates (latitude, longitude) with elevation, slope, mean annual precipitation, mean annual temperature, and cation-exchange capacity data. To achieve this, we extract the information from look-up-tables (GeoTIFFs, georeferenced images) that are taken from Google Earth Engine (Gorelick et al., 2017) and `soildgrids.org` (Hengl et al., 2017).
Expanding this framework to include new datasets provided as GeoTIFFs is very smooth. For more details please refer to the `readme.md` present in our git repository[1].

3.3.2 Soil Data

The GIZ and its local partners do not conduct soil measurements on the farms participating in surveys: it is impossible to use the bare soil measurements directly. Because soil information is reportedly a key yield predictor, we attempted to build a system to infer soil measurements at the locations of interest, i.e., farm locations. This system would have enabled to obtain approximate soil data using only geographical information and thus eliminating the need for costly laboratory analysis.

We used a Gaussian Process Regressor (also known as universal Kriging) and trained for each soil variable we wanted to infer: we did not use the other variables to predict. We used Leave-one-out cross-validation and observed no significant correlation between our predicted soil values and the values from the dataset. *We thus conclude that it is not possible to capture the soil distribution with the samples at hand or that one cannot infer the soil properties solely based on spatial coordinates.*

We also attempted using the data provided on `soildgrids.org`. However, a comparison of their data with the lab results provided by the GIZ showed no correlation. Hence we recommend cautious use of this data source without testing data to estimate its accuracy.

[1] https://gitlab.com/analytics-club/hack4good/hack4good-fall-2021/giz-soil/giz-soil/

4 Results

4.1 Recommendation engine

After analyzing the data, we concluded that a recommendation engine, a system that mimics the expert's decision on which farming practice to apply given a farm, is impossible with existing data. Currently, data is collected after the farming practices are recommended and then applied. Using this data would create a model that guesses which farming practice is currently applied, not which practice the expert would recommend. This system is not valuable to the GIZ - we did not build it. Instead, we provide possible improvements for future data collection workflows in section 5.

4.2 Yield prediction

Two surveys at our disposition contained yield data, the Yield Surveys, and Adoption Surveys. The latter also contains precious features to predict the yield, such as household information or farm area. However, the yield values it contains are self-reported by the farmers in units of bags, and hence we assumed it is less precise.

4.2.1 Yield comparison between the adoption surveys and the yield datasets

To choose which dataset to use, we compared the distribution of yield values across the two datasets. Though they are not in the same units, the distributions should be identical up to unit conversion if there is no underlying bias.

Figure 1: Comparison of beans yield distribution in the two surveys

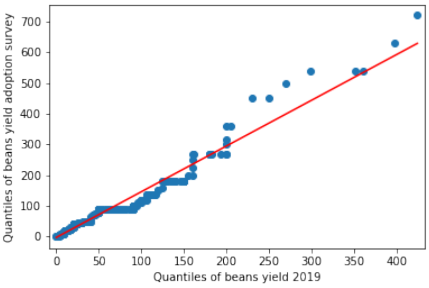
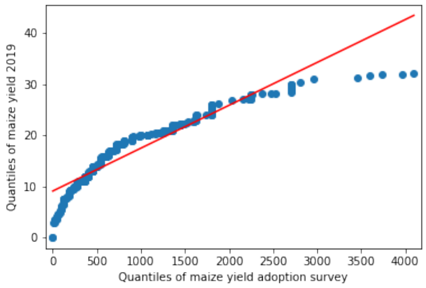


Figure 2: Comparison of maize yield distribution in the two surveys



The plots in figures 1 and 2 (Q-Q plots) compare the two sample distributions. Crucially, if the distribution are the same up to a unit-conversion then *we should observe a linear relationship between the two distributions: the points should follow a straight line going through 0.*: it occurs for the beans but not the maize distributions. Because of this, we believe there is a bias in the underlying data, the farms interrogated in the two surveys seem to differ. For example, it could be that the average farm is bigger in one of the surveys. Consequently, we chose to use the Yield Surveys with scientific measurements to predict yield as we could not explain the bias.

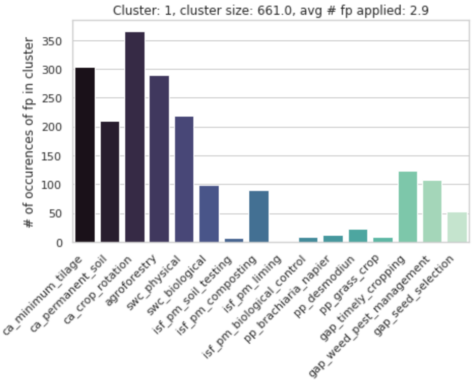
We trained our prediction model on the yield data from 2019 and 2018. The main features of this model were farming practices applied and inferred soil information (which we later realized was of inferior quality). In particular, we did not have land area nor household size features as these essential features were not present in the yield data set. Moreover, there was uncertainty about the units used for the yield features. These confounding factors lead to unsatisfactory results whatever the model (Lasso, ElasticNET, random forest, gradient boosting), with an R^2 error of around 0.2 for each. Therefore, we did not analyze the models as our results would not have been meaningful. Instead, we collected recommendations for the GIZ. We list them in section 5.

4.3 Clustering of farming practices

The GIZ recommends practices in groups, but these are not applied together. To investigate this issue, we first analyzed the frequency of combinations of farming practices (e.g., only practice A and B) among our data. We ran a clustering algorithm on the data containing only the farming practice features with one row for every farm. Figure 3 contains the count of occurrences of each farming practice in one of the clusters.

With 661 farms, this cluster contains more than a third of the data we had at our disposal. Secondly, farms in this cluster apply only a few farming practices: 2.9 on average and a median of 3 compared with 6 overall and up to 10/12 in other clusters.

Figure 3: Clustering results for farms with few farming practices



There seems to be a pattern where farms that implement few farming practices (their feature vector are close to 0) choose mainly a combination of the ones which have high bars in the graph of figure 3. In other terms *farmers who implement only a few farming practices prioritize those over the rest.*

5 Recommendations to the GIZ

Our recommendations to the team are the following:

- For columns containing many unique categorical values (e.g., fertilizer), it is important to group the values in more general categories to use them in a model. Hence, we recommend either collecting less granular data or, better: providing a grouping of all the categorical values.

- Standardizing column names and option names in multiple-choice questions shared over surveys would allow for quicker analysis and remove tedious work. For example, the adoption survey used numbering for Integrated Soil Fertility and other measures, yet the 2019 yield data used a different notation.

- The GIZ should use standardized units when collecting data. Identical features across surveys should use the same unit and specify those in the feature name.

- The GIZ should implement a standardized approach to record features containing multiple values. For example, when writing down the fertilizers used or crop variety information in the same cell, the surveys should use consistent separators (e.g., commas, semi-columns, or space) and naming.

- Discrete data is valuable. However, many related binary features reduce readability and delays preliminary analysis and processing. For example, a single column combining all farming practices applied (with standardized separation and naming) and thus combining all related features into one. Similarly, associated measures, animals on the farm, etc., can also be combined into only one attribute.

- There are inconsistencies in the data and these should be investigated if possible. Notably, in some cases, a group of farming practices is applied when none of the sub-practices are.

- The GIZ should adopt a baseline of features to collect in all surveys, with standardized names and units. We recommend yield, land area, available workforce, household information, fertilizer application, inter-cropping varieties, years of operation, and soil testing information if possible. It would be possible to keep core features even when merging datasets with these.

- The adoption of an identifier for farmers, consistent across surveys, would allow tracking the same household across surveys. This identifier would enable richer analysis by combining the results of surveys. These can be designed to preserve anonymity.

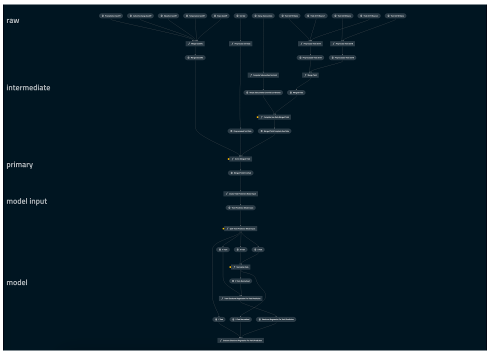
- Development of a recommendation engine would require data acquisition at the first visit. This measure could also allow for a diversification of the data sets. Current data only contains farms affected by the GIZ program and applying farming practices, collecting data at first contact would remove that bias. The complete view of the distribution would allow better analysis and more generalized models.

6 Deliverables

6.1 A Complete Data-Science Pipeline

To ensure our results would be easily reproducible and extendable, we used Kedro (Bălan et al., 2020) extensively. The framework provides a structured way to reason about data pipelines and their building blocks. We provide extensive documentation for the code along with installation instructions in the `readme.md` in the project's Git.

Figure 4: Visualisation of the yield prediction pipeline



Please note that the datasets provided by the GIZ contained sensitive information and are thus not published on Git.

6.2 Data enrichment process

To complete our data, we developed a process that enables us to add information to data points with WGS84 information as described in 3.3.1. The existing code can easily be extended to enrich with more data contained in GeoTIFF. Please refer to the section "Extending the enriching" in the project readme in our git project for more details [2].

7 Future work

As the project was only eight weeks long, we could not pursue multiple interesting areas of improvement or analysis. On top of the recommendations, we list areas we believe are most impactful to improve the project further.

- Soil features inference could benefit from using more input variables such as the elevation, slope, mean annual temperature, and mean annual precipitation. All of which can be obtained by the framework described in 3.3.1. There seems to be some scientific literature on the subject.
- Identify sources for distribution disparity described in section 4.2.1. Without detailed investigation, this diparity could risk corrupting results obtained with the Adoption Survey.
- Deep dive in the scientific literature regarding yield prediction. The pitfalls we encountered during the project might be documented in related papers. Because of time constraints, we could not do it within the project's time frame.
- Develop a process to collect standardized data with well-defined categories and columns. It would be possible to create automated analytics dashboards (e.g., Tableau / Microsoft BI) to always have up-to-date metrics on the program's impact.

References

Lorena Bălan, Kiyohito Kunii (Kiyo), Dmitrii Deriabin, Lim Hoang, Andrii Ivaniuk, Yetunde Dada, Deepyaman Datta, Zain Patel, Gordon Wrigley, Ivan Danov, Jo Stichbury, Nasef Khan, Nikolaos Tsaousis, Merel Theisen, Waylon Walker, Tam Nguyen, Richard Westenra, Lais Carvalho, Marcelo Duarte Trevisani, Sebastian Bertoli, Shahil Mawjee, sasaki takeru, Bas Nijholt, Dmitry Vukolov, Kody Fischer, Vijaykumar, Yusuke Minami, bru5, and dr3s. 2020. quantumblacklabs/kedro: 0.17.0.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Tomislav Hengl, Jorge Mendes de Jesus, Gerard BM Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, et al. 2017. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2):e0169748.

[2] https://gitlab.com/analytics-club/hack4good/hack4good-fall-2021/giz-soil/giz-soil/