

Assessing internal and external drivers of land productivity in Eastern Zambia

Frederike Lübeck, Christoph Mony, Kaoru Schwarzenegger, Vincent Bardenhagen

Abstract—This project presents a methodology for modeling crop yield variation in Eastern Zambia (EZ) based upon publicly accessible soil and weather data. For this purpose, we employed survey data from the "Gesellschaft für Internationale Zusammenarbeit" (GIZ), an NGO, to access socioeconomic data for EZ and land productivity for two value chain crops: groundnuts and soybeans. The final data pipeline allows to combine various data sources to learn about the distribution of soil properties (e.g., sand proportion), meteorological variables (e.g., droughts), or location-specific demographic data (e.g., access to agricultural knowledge). We show that modeling land productivity with modern machine learning algorithms proves to be difficult due to large uncertainties in the data collection process. Nevertheless, we can provide value by making the prepared soil and weather data easily accessible for use in combination with local expert knowledge to allow for the issuance of early warnings.

Keywords—Food stability, Land productivity, Machine Learning, Geospatial analysis



1 Introduction

Agricultural production is vital to most Sub-Saharan countries like Zambia, employing 52% of the population [1]. Nevertheless, 34% of all children suffer from impaired growth and development induced by malnutrition [2]. Also, in the future, the food demand in these areas is expected to increase dramatically, and small to medium scale farmers will likely play a crucial role in sustaining this growth [3].

The agricultural output in Sub-Saharan farms strongly depends on rainfall variability and temperature extremes [3]. Flash floods and droughts are frequent events destroying large parts of the harvest. Moreover, in some areas, soil properties might be less suitable for specific crops, negatively impacting the yield. For this reason, we set out to identify specific drivers in soil and weather data explaining variation in crop yield for the two crops promoted by GIZ: groundnuts and soybeans. Such knowledge could inform mitigation strategies against detrimental meteorological events and could guide future plantation decisions.

The rest of this paper is structured as follows: Section 2 gives an overview of the employed methodology. Section 3 reviews the data sources used to build the final dataset. Section 4 presents the algorithms utilized during the statistical modeling. In section 5, we assess the performance of the models and discuss the quality of data.

2 Methodology

The developed methodology encompasses all crucial steps of an analysis pipeline from data acquisition, cleansing and merging of data sources, feature engineering, visual analysis, and machine learning. First, the data acquisition step retrieves relevant soil and meteorological data from online APIs and transforms them into tabular format. Second, the cleansing of data sources mainly refers to the GIZ surveys, which were made temporally consistent. Furthermore, the survey's location information allowed us to merge survey, soil, and weather data spatially. Third, during the feature engineering, we extracted specific information such as flood and drought occurrences for further analysis. Lastly, we trained several machine learning models on the final dataset, intending to identify the most important drivers of land productivity.

3 Datasets

This study was comprised of three different data sources, *survey data*, *soil data*, and *weather data*, that are described below. The final dataset (exemplary variables shown in Table 1) used in our analysis was created by mapping the soil and weather data spatially and temporally to the GPS locations of the villages in the surveys (see Figure 1).

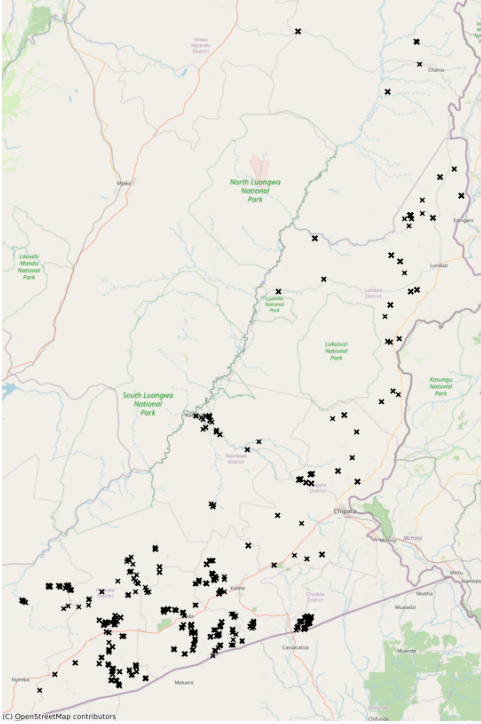


Figure 1: Area of study in EZ. Black crosses denote village locations surveyed by GIZ.

3.1 Survey data

We obtained the productivity data, which served as our target variable, and other socioeconomic variables in the form of GIZ surveys from 2016 to 2019. In each year, about 500 random farmers within Eastern Zambia were interviewed about different agricultural aspects. A major challenge arose from a varying set of questions and variable names between the years. Furthermore, about half of all surveys could not be associated with any GPS location and were not considered for the analysis. The remainder of the data was cleansed manually and merged into one dataset, consisting solely of the features that appeared in all survey years. The land productivity LP was calculated for both groundnut and soy harvests

$$LP = \frac{M}{A},$$

where M denotes the harvested crop weight (in tons) per year and A the field size (in ha) dedicated for the plantation.

3.2 Soil data

We retrieved soil data for various variables such as sand content, pH-value, or nitrogen amount from SoilGrids¹. The data originates from globally fitted

Table 1: Exemplary variables in our analysis.

Variable	Units
Survey data	
Crop harvest	tons
Conducted number of trainings	counts
Soil data	
Soil pH	pH
Proportion of sand particles	g/kg
Weather data	
Start of rain season	day of year
Total rain season precipitation	mm
Flood events (per season)	counts
Monthly mean temperature (Oct–Apr)	K

models and incorporates measurements taken all over the world [4]. Especially in regions like EZ, where not many measurements or local models exist, this data allows for coarse first insight. The data had a resolution of 250 meters, and at each grid point, one could investigate the soil properties in various depths (e.g., 60–100 cm). For our purpose, we accessed all variables in EZ over the public WCS API in Python.

3.3 Weather data

For the weather data, we relied on the most current reanalysis ERA-5 conducted by the ECMWF [5]. More precisely, we used the ERA5-land reanalysis, which implements atmospheric fields with a spatial resolution of approximately nine kilometers and a temporal resolution of one hour. We accessed the most relevant features such as precipitation and temperature using the public CDS API² in Python. Afterward, we used the data to engineer more features like drought or flood occurrences.

4 Machine Learning

Our resulting dataset gets split into a training (70%) and a test set (30%). Since our total dataset included about 430 samples for groundnuts and 235 samples for soybeans (too little for effective cross-validation), we performed a random search over various hyperparameter configurations by fitting XG-Boost, CatBoost, Random Forest, Elastic Net, and Ridge Regression models to the training set. Afterward, we evaluated the performance on the test set using the R^2 and RMSE metrics.

1. <https://soilgrids.org/>

2. <https://cds.climate.copernicus.eu/toolbox/doc/api.html>

5 Results

The results of the modeling step showed to be rather unsatisfactory, yielding with an R^2 score of 0.14 for groundnuts and 0.08 for soybeans only slightly better results than predicting the average. These results were the best values obtained after testing different models with various hyperparameter specifications.

Further inspection of the land productivity for the different villages revealed considerable variation, even within single villages. Soil and weather variables are constant at each village since the survey data only included the closest village and not the exact farm location. Hence, they cannot explain this internal variation. Moreover, even after including more variables to control for socioeconomic differences within villages, the results did not improve substantially. Three potential explanations for this limited ability to model the data were identified as

- 1) Data collection in the field is a challenging task. Especially for our target variable, the units of M and A were numerous, and the resulting values are more often rough estimations than precise measurements.
- 2) Inaccuracies in GPS measurements. We could only rely on GPS information of the closest village where the survey was conducted, but not of the farms themselves.
- 3) The information about the farmer's characteristics and practices, retrieved from the surveys, contained too few explanatory factors. This might be due to the fact that the responses to many questions did not show any variation.

However, even though we could not link soil differences or meteorological variations to land productivity via the survey data, in discussions with the experts from GIZ and Comaco, we learned that the soil and weather data we used has been unknown before. For this reason, we decided to focus on making this data available in an easily accessible manner to the local people in charge in the form of a Tableau dashboard³. This dashboard enables them to combine this data with their expert knowledge and hopefully draw actionable conclusions.

6 Conclusion

The data-driven identification of drivers in agricultural productivity is a challenging task. In prin-

ciple, this should be possible when including future well-designed surveys, which allow for more accurate and homogeneous land productivity data. Until then, we hope that consistently analyzing external factors, such as soil properties and weather data, helps to design prospective interventions.

Therefore, we visualized these data sources on a dashboard that can be used to investigate regional and temporal variations in these external factors. This way, soil properties and weather data are made easily accessible, and we believe that combination with expert knowledge allows for the identification of causal effects and driving factors.

References

- [1] Rogan, J., 2015. Zambia Country Analysis. *UN Reports Zambia*. Online access via http://zm.one.un.org/sites/default/files/un_country_analysis_report.pdf.
- [2] Zambia Statistics Agency, Ministry of Health (MOH) Zambia, and ICF, 2019. Zambia Demographic Health Survey Summary Report 2018. Online access via <http://www.dhsprogram.com/pubs/pdf/FR361/FR361.pdf>.
- [3] Zhao, Y. et. al., 2018. Comparing empirical and survey-based yield forecasts in a dryland agro-ecosystem. *Agricultural and Forest Meteorology*. Volume 262. Pages 147-156. Online access via <https://doi.org/10.1016/j.agrformet.2018.06.024>.
- [4] Hengl, T. et. al., 2017. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS One*. Volume 12:2. Pages 1-40. Online access via <https://doi.org/10.1371/journal.pone.0169748>.
- [5] Hersbach, H. et. al., 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. Online access via <https://doi.org/10.1002/qj.3803>.

3. Accessible online via: https://public.tableau.com/profile/chris3097#!/vizhome/GIZ_16056548246950/Story1?publish=yes