Predicting farmers yields in rural Kenya

Giuliana Daga

12/11/2020

Contents

Introduction

This report analyzes spatial aggregation and potential sources of improvement of Bima Pima, an index-based microinsurance product offered in Siaya County, Kenya. The product is piloted in Kenya by Agriculture and Climate Risk Enterprise Ltd. (ACRE), operating in Kenya since 2009 and now works in over ten countries across Africa and South Asia. An index-based product differs from more conventional insurance in that it links payouts, not to actual crop losses but exogenous events. In this case, ACRE takes measured rainfall, which likely correlates with crop losses. This way of assessing risk offers many advantages. It eliminates the need for expensive crop evaluations, allows for rapid payouts, and eliminates moral hazard, given that farmers have an incentive to make their farms succeed. Despite such promising advantages, basis risk, defined as the correlation between insurable events and yields, remains an issue. This study aims to improve this imperfect correlation, especially at the low end of the yield distribution.

The purpose of this report is twofold. On the one hand, we will transform statistical analysis into a geospatial one with useful visualizations that improve our understanding of the targetted farmers and their characteristics. For this purpose, we will use survey data to map farmers at the center of their farms. We will then overlay this data with the Kenya counties' shapefile and a fishnet based on ACRE's payout aggregation and calculations to observe their characteristics. On the other hand, we will use machine learning methods to explore whether different weather data help us predict farmers' self-reported profits in the Short Rains season 2018-2019. Given the extreme difficulty in this challenge, we will succeed if we learn whether adding different weather variables improves our understanding of farmers' performance to protect smallholder farmers facing considerable risks.

This report will proceed as follows. We will first state smallholder farmers' vulnerability to weather hazards and discuss weather index insurance (WII) as a development solution. In the following section, we will describe our weather data sources, their availability, and levels of granularity, as well as a brief discussion about the process to merge them and the unit of analysis. The analysis section describes our data and the methodologies we chose for our data. The results section comments on the metrics obtained through the supervised machine learning regression models. To conclude, I will discuss the successes and pitfalls of my project and identify opportunities to move forward in the future.

Problem Statement and Background

Smallholder farmers are increasingly affected by weather forces beyond their control, which affects their yields, profits, and livelihoods significantly. Higher crop yield variability may increase poverty and food insecurity, especially in developing countries (Wheeler, 2013; Bown, 2015). However, less than 20 percent globally have insurance coverage to protect themselves against shocks (GSMA, 2020). After catastrophic weather events, smallholder households may end up selling assets to smooth consumption (Carter, 2006).

On the demand side, low awareness and willingness to pay, together with liquidity constraints, lousy perception of the product (a lack of trust in the business), and behavioral biases such as present preferences (farmers prefer payouts closet to the current time). On the supply side, insurance providers have historically overlooked the market, given that the higher cost of serving rural customers makes it a less profitable segment of the industry.

Weather index insurance (WII) could potentially overcome some of the problems with traditional insurance schemes. Through this method, insurance providers calculate payouts subject to objectively measured weather data, which presumably correlates with production losses. Compared with conventional insurance products, WII imply fewer administration costs, making them more affordable and giving faster payments to farmers (Sibiko, Veettil, & Qaim, 2018).

However, despite all these advantages, smallholder farmers haven't voluntary uptake the product as anticipated. Some authors think that adding to the previously mentioned challenges for farmers (liquidity constraints, preference biases, etc.), the basis risk that often remains is a significant issue (Norton, Turvey, Osgood, 2012). Flawed basis risk means that the index insurance's construction may not match the insured's risk exposure. Index insurance services have been in the industry for over the last ten years, using mobile technology to digitize farmer's registration, premiums, payout claims, and satellite technologies to assess service delivery. This report will explore the basis risk's local dimension, looking for satellite weather data that best describes the index/ yield relationship for the Short rains season 2018 in Siaya County, Kenya.

Data and wrangling methodology

The data in this project is composed using three primary data sources:

- **Georgetown Initiative on Innovation, Development, and Evaluation (Gui2de)** - Household survey data provides a measure of, among other things, maize's harvest and farmer profits on Baseline (before first planting season, Jul-Aug 2018), Midline (at the end of the first planting season, Feb-Mar 2019), and Endline (at the end of the second planting season, Nov-Dec 2019). For this analysis, we will focus on the first season; therefore, we will use Midline.

- **IRI/LDEO Climate Data Library** is a library of datasets. We will use the NOAA NCEP CPC CAMS, with daily precipitation data measured in milliliters for each pixel. There is availability from 1983 up to date. However, we will only use it from September 2018 to January 2019.

- **NASA Giovanni (Geospatial Interactive Online Visualization ANd aNalysis Infrastructure) data collection**: obtained daily data such as atmospheric pressure, atmospheric temperature, surface thermal properties, soil temperature, wind speed, radiation, etc.

The unit of analysis for this project differs between the dependent variable and our predictors. We have data for the short-rains season 2018-2019 at the individual level for the dependent variable. For the predictors, we got daily data disaggregated at the pixel level, as defined by ACRE by the first decimal in latitude and longitude (i.e., 34.3-0.2; 34.3-0.3). However, to make more precise predictions, I aggregated daily weather data into four defined periods in the planting season: germination, vegetation, flowering, and pre-harvest. I included their mean, sd, and max for each period.

Our sources of data have some pitfalls: 1. Nearly 40% (800) of our farmers don't farm maize in the short-rain season and prefer to wait for the long rains season. 2. There is a potential self-reporting bias that we will discuss later in this paper. 3. Obtaining weather data for our particular pixels was very challenging, and we hope to improve this process in the future and cross-validating our metrics across sources.

One of the challenging parts of this project was to join all datasets as we want them. Survey data's unit of analysis was at the household level, and it included latitude and longitude for each of the smallholder farmers. We downloaded a shapefile containing county divisions in Kenya. We built a grid (or fishnet) around each of the dots provided by ACRE to localize our farmers in a broader context. We used the "sf" package extensively to join, filter, and overlay different shapes together. For the rest of the analysis, we had to ensure that all spatial features overlay one with the other. More specifically, making sure our farmer grid, survey, and weather data aggregate farmers at the pixel level correctly.

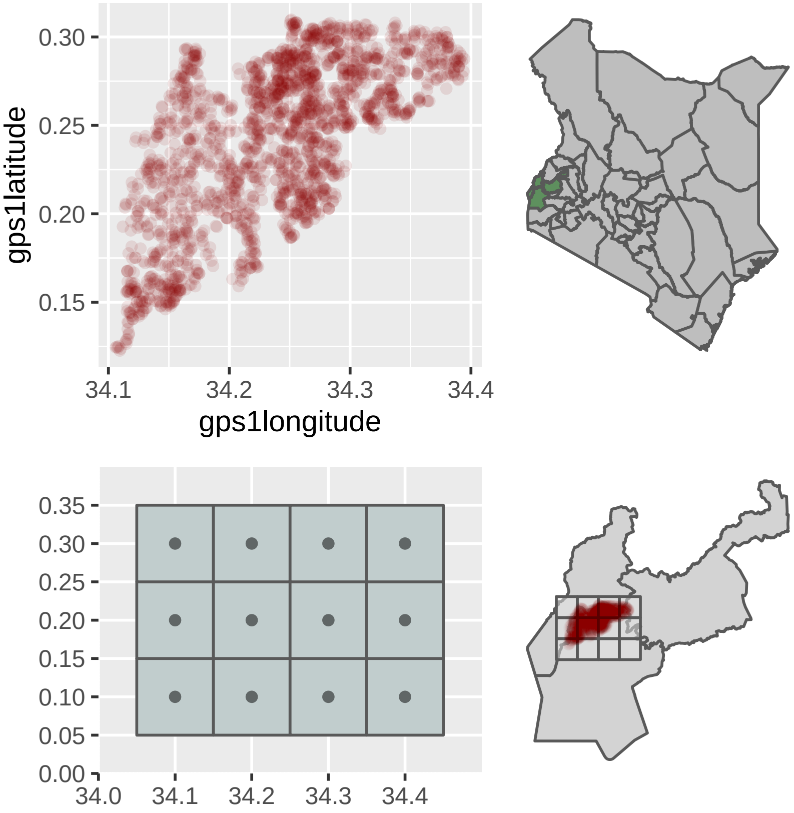

Figure 1: Farmers' location in broader Kenya

Once I got the .csv files, I had to do a lot of data wrangling and manipulation. First of all, I had to read each dataset with daily rainfall data for a given pixel inside a loop. Then, I converted the "time" variable into a date variable and restricted the sample to the season under consideration (September 2018 to January 2019). I also created a variable indicating each of the four periods: germination, vegetation, flowering, and pre-harvest. It is important to note that these periods usually start from different windows planting windows among farmers. In this case, we imputed the same planting date for all farmers in the sample to assign our periods. After creating the periods, we collapsed at the period level, getting the mean, sd, and maximum values of each weather variable. We then transposed the data as a matrix, renamed columns and rows, and pivot wider so that each row is a pixel code and each column a different variable and metric. Finally, I fully joined our survey data with our weather data by the "pixel_code." I also had to make sure that I was doing the spatial join correctly with the "st_join" function.

The full dataset was analyzed for missingness and the potential for imputation. We found that we had daily data for all pixels included in our analysis. However, out of the initial 2,210 farmers, only 1,313 remained.

Analysis

We start the analysis by excluding those observations with missing values in the variable "Kg. of maize per hectare". This procedure leaves us with 1,282 final farmers with self-reported maize productivity.

We can see a wide variation in the levels of productivity, investment, overall profits, and payouts. We also graph some of our predictors' mean to see how they look and spot any variation. The same happens with our rainfall and other weather data, like humidity:

Before running our predictive models, I set seed and split our data into a training and test dataset to conduct the analysis, partitioning 75% of the former and 25% to the latter. Our strata for the split is kg. of maize per hectare. We then build a recipe to: impute missing values, which is less than four for all variables; take the logarithm of our dependent variable,

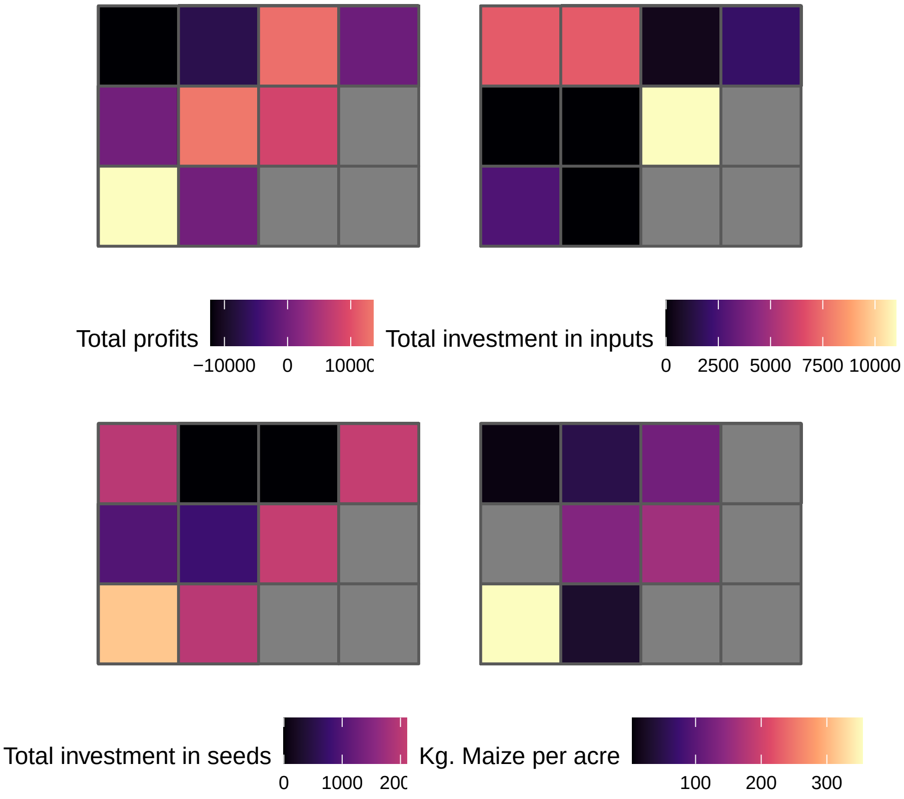

Figure 2: Investments and productivity at each pixel

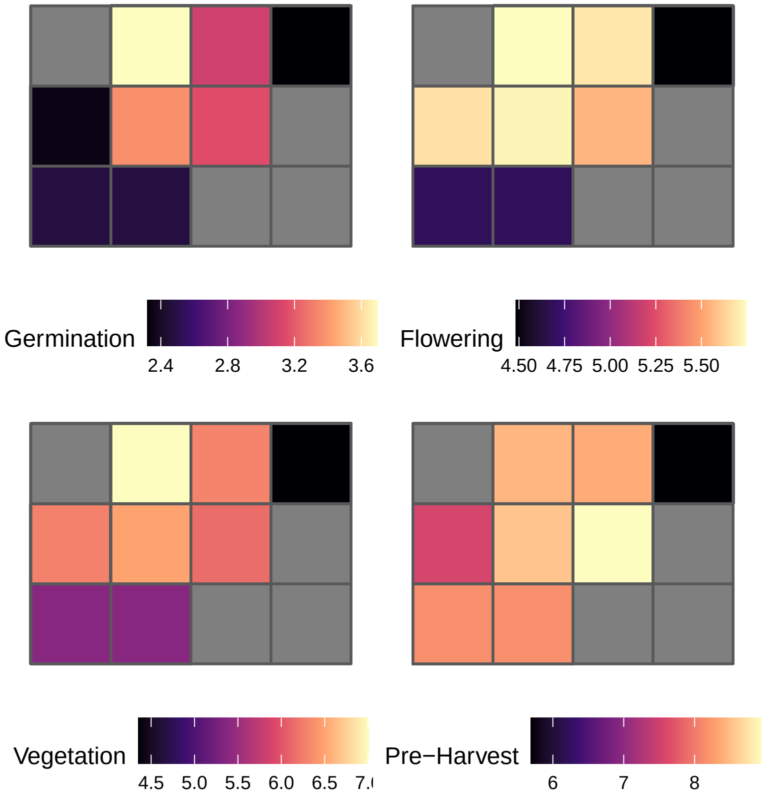

Figure 3: Rainfall data at each pixel

off-set by a factor of one, and normalize the scale of all our numeric variables. I then set seed again to partition the data into five-folds, to be able to set up a k-fold cross/validation procedure, using our dependent variable as a stratum. Setting up this control condition will allow us to validate our results across subsets in our data, preventing the model from overfitting the data.

I explored three supervised learning regression models to predict farmer's yields. I tested: k-nearest neighbors, regression trees, random forest, and support vector machines with Polynomial Kernel. Given our data's non-linear nature, it is unlikely that the linear regression model would provide a good fit. We also tried k-nearest neighbors, one of the simplest machine learning methods with few parameters to tune. The drawback is that it doesn't support automatic feature interaction, as the regression trees and random forest. We think that a regression tree model would perform better, given that it is more robust to noise and outliers relative to the LM model. I suspect that the random forest model will achieve the best results given that it combines many decision trees into a single model. However, given that each of these four models entails different strengths and weaknesses, I test each on the training dataset to determine which model has the most robust predictive performance.

Results

Unfortunately, none of our four models predict changes in farmer's yields satisfactorily. If we follow the Root Mean Square Error (RMSE) approach, we will find that even though all models perform similarly, the KNN slightly smaller with an RMSE of 0.17. We can also see that KNN's has a somewhat higher r-squared of 0.047. These results are not as high as we would like to but are higher than the industry benchmark (currently around 0.02). I followed many strategies to improve our model. I tried to expand the gris allowing trees to go deeper, but the the little correlation between variables didn't allow my code to run. I also tried different dependent variables (total yields, productivity per hectare, real profits), and log transformed the dependent and independent variables.

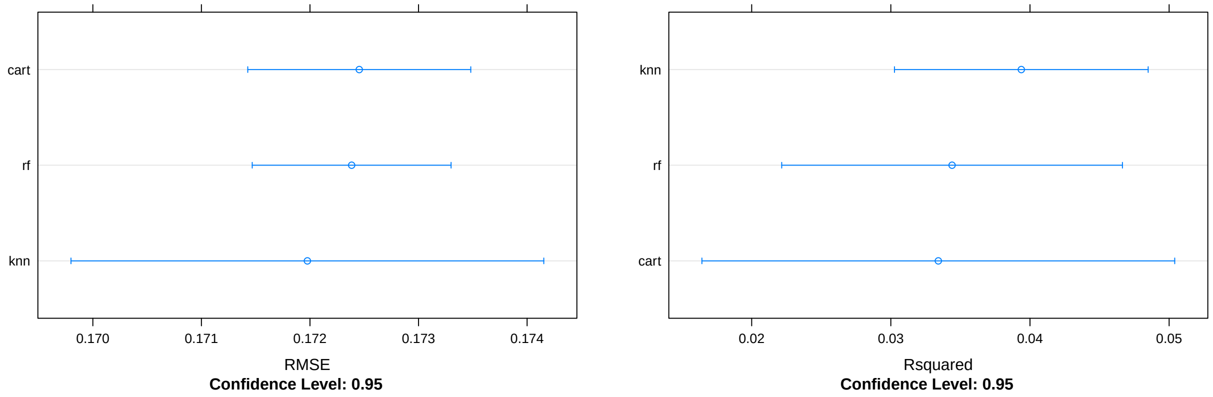

Figure 4: Machine learning model comparisons

Regarding variable importance, different models selected different variables. For example, the KNN didn't choose any variable over the other, whereas the regression tree assigned 0.065 importance to humidity in the germination period. Lastly, we can see how the model assigns importance to some variables for the random forest model but their confidence interval for most crosses cero. However, variables la the net shortwave radiation maximun in the flowering period and surface runoff vegetation mean matter.

Discussion and conclusion

I based the success of this project on two broader objectives. First, to translate purely statistical analysis to spatial data analysis. In this matter, we could visualize the high concentration of farmers in a specific area of Siaya county. Moreover, we observe the spatial aggregation of farmers' yields, payouts, and weather indicators more clearly. On the other hand, we wanted to see if throwing extra weather data into a machine learning exercise would improve our farmer yield predictions to protect them better against weather hazards. Even though we improved the industry's benchmark, we could not predict a high percentage of farmer yield variation in this particular setting.

There are a variety of reasons that explain our not satisfactory results. Reduced spatial variation (high concentration of farmers) and time variation (only one season) represent

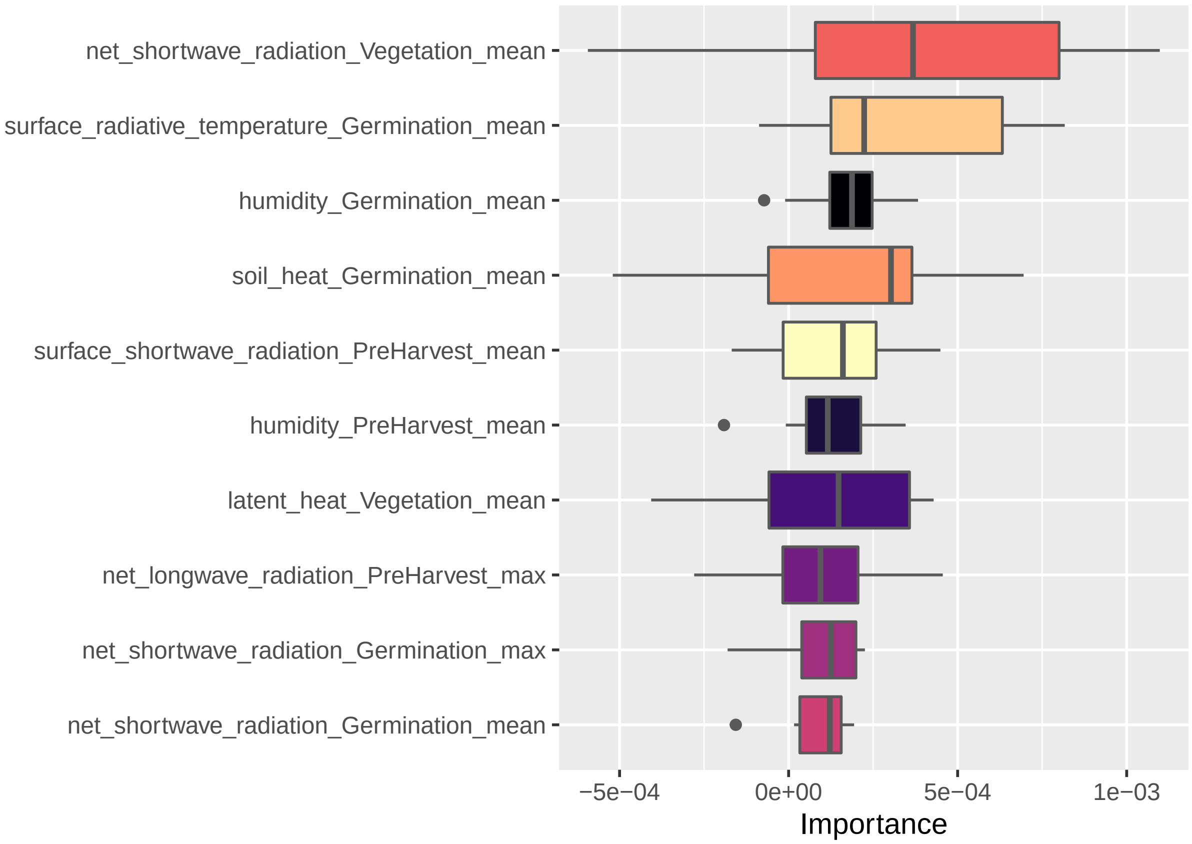

Figure 5: : Random Forest: Top 10 Variables of Importance

enormous challenges for this type of analysis. Even though we spotted this concern at the early stages of the study, we decided to move forward, given that this assignment is embedded in a wider project, including a Randomized Control Trial (RCT) that included only the selected farmers. Another challenge is the nonrandom measurement error in self-reported yield statistics, which can threaten external validity concerns. Potential biases and measurement error are well-documented in the literature, and we think satellite crop images can account for this fact.

I see many opportunities for improvement. I want to extend this analysis to the historical (20 to 80-year) weather and yield data to develop a weather index-based insurance product in future endeavors. Moreover, it could be useful to build and test our index using non-parametric and quantile regression to better protect those at the lower end of the distribution.

Apendices

a. Variable description:

Variable	Description
Maize productivity	Self-reported kg. of maize per hectare
Rainfall	Precipitation measured in milliliters
Humidity	Water vapor contained in the atmosphere
Evapotranspiration	The sum of evaporation and plant transpiration
soil heat	The rate at which heat is transferred through
Soil moisture	The depth-averaged amount of water in a soil layer
Surface air temperature	The average temperature of the air
Surface pressure	The atmospheric pressure at the Earth surface
Surface wind speed	Air movement speed relative to a fixed point
Surface heat	The average temperature of the Earth's surface
Latent heat	The rate at which heat is transferred through a given surface.
Net long/shortwave radiation	Difference between incoming and outgoing radiation
Surface radiative temperature	Average temperature of the Earth's surface
Surface runoff	Water, from rain, snowmelt, etc. which flows over the land surface.

References

Gaigné, C., Christopher B. Barrett and A. Mude. “Index Insurance Quality and Basis Risk: Evidence from Northern Kenya.” Development Economics: Regional & Country Studies eJournal (2016): n. pag.

Norton MT, Turvey C, Osgood D. Quantifying spatial basis risk for weather index insurance. J Risk Finance. 2013;14(1):20–34.