Predicating Food Insecurity in Sub-Saharan Africa with Machine Learning

Yujun Zhou

April 13, 2018

Research Question

- Can we build an early warning system of food security in areas where data is scarce and data collection is costly ? (Hutchinson,1991)
    - Timely and accurate targeting is essential for aid and humanitarian responses
- How to make use of publicly available and economically meaningful data?
    - Price data of the main agricultural markets are collected monthly or weekly
    - Precipitation/temperature/soil quality from satellite imagery are relevant to agricultural production
- Find the supervised learning approach that has higher predicative power but remain interpretable.

Preview of Results

- Predictions from our model explains 50%-70% of cluster level variation and the result is consistent across three different countries in different years.
- Validation of "A Prototype for Predicting Food Insecurity Using Readily Available Data" paper with three countries and seven years of training data.
- Using the same sets of features, a tuned machine learning model outperforms a baseline linear model by xylem %.

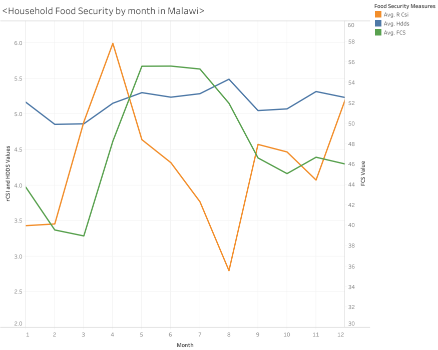
Literature Review

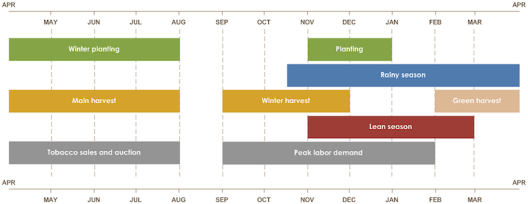
- Night lights data (Chen and Nordhaus 2011;Henderson et al. 2012.) does a good job at predicating economic activity but the variation is little in areas of the extreme poor or in urban areas.
- Mobile phone data (Blumenstock et al.,2015; Steele et al.,2017) is useful but expensive.
- High resolution satellite imagery are becoming cheaper but highly unstructured and contains measurement error (Jean et al., 2016; Donaldson and Storeygard, 2016).
- Convolutional Neural Network (CNN) models (Jean et al., 2016; Babenko et al. 2017) can explain an average of 46% of the variation at village level but they require an enormous amount of training data and are computational extensive. Interpret-ability and repeat-ability are not that promising (Engstrom,2018).

Framework

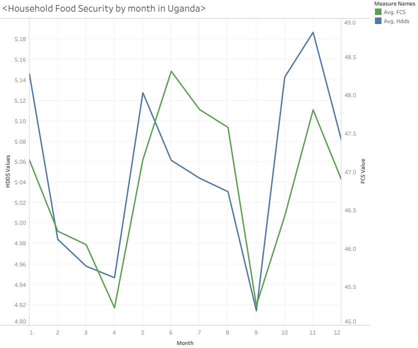
- Understanding of the food security. The definition of food security has various characteristics. The use metrics can lead to quite divergent rankings of the same population (Steele et al. 2017)
- Geo-referenced household surveys (LSMS data) allow us to explore the spatial-temporal variations in food security measures.
- The sampling framework in these surveys made it possible for us to observe a nationally representative sample in different months and agroecological zones.
- Explain these variations by the spatial-temporal variation in food availability and in food access.
  - Align weather data to crop growing season
  - Align households to the most relevant market price

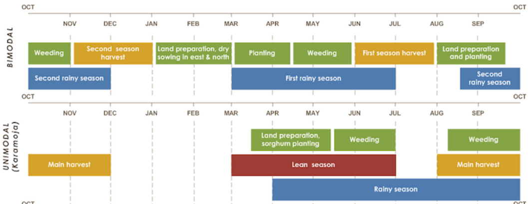
Temporal Variation





Temporal Variation





Spatio-temporal variation

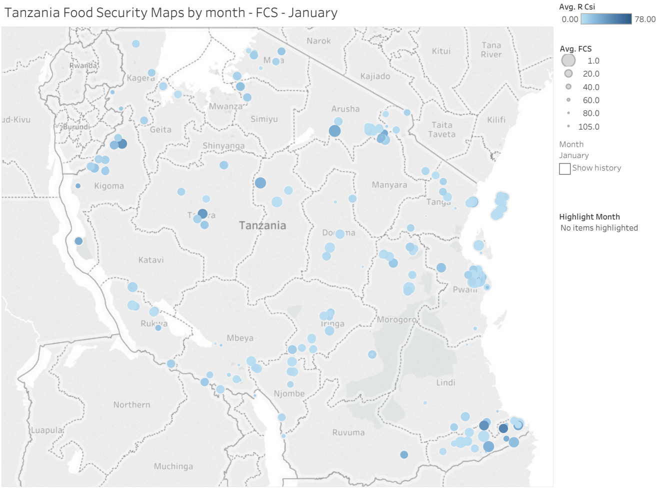


Figure 1: Food Security Maps in Tanzania, January

Spatialtemporal variation

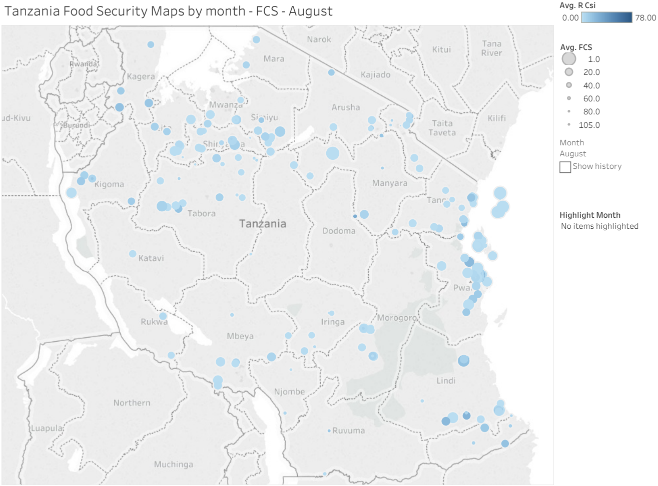


Figure 2: Food Security Maps in Tanzania, August

Discussion on the types of regression models

Data

- Price data (maize and cassava)
- Weather
- first day of rainfall
- rainfall in growing season
- maximum of days without rain
- temperature: growing degree days
- Soil quality: Slope, water, soil nutrition
- Roof/floor type, household asset index

Modelling strategy

- sets of models deal with high dimensional problem with regularization:
  - Lasso
  - Ridge
  - Elastic Net
- models that are great with heterogeneous distribution of data:
  - Random Forest
  - Bagging / boosting

Main Results

- scatter plots

Future Steps

- Vary the time gap between training and testing (train and test on the a subset of data that are only several weeks/month apart)
- Trained on a pooled data set across different countries V.S. Fit models on each individual country with the same procedure
- Predict "now": countries/areas that are not surveyed and suggest areas that are likely to have a food shortage.

Limitations

- Gradual food insecure compared to sudden, abrupt threat to food security (natural disaster, war and conflict)
-