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)
- Can we 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
- Can supervised learning methods have higher predicative power than linear models ?

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 several years of data
- Using the same type of variables, a tuned machine learning model outperforms a baseline linear model by xylem %.

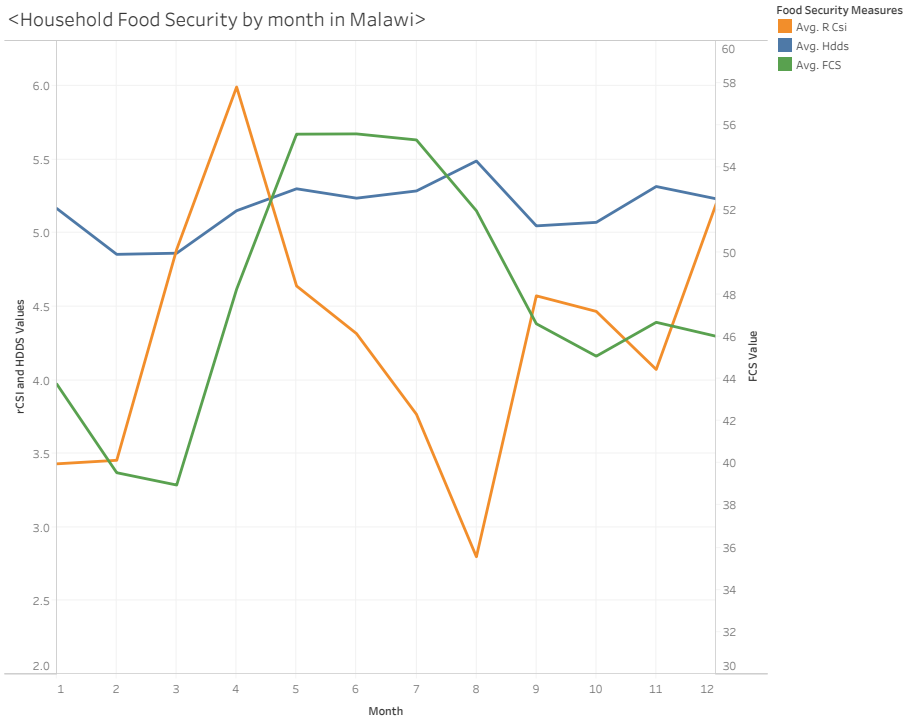
Literature Review

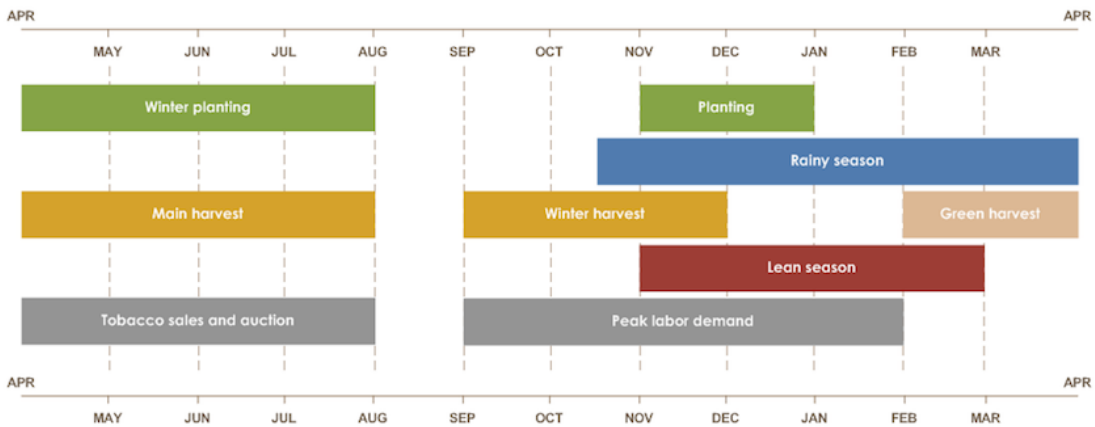
- Night lights data (Jean et al., 2016) is noisy and variation is little in areas of the extreme poor
- Mobile phone data (Blumenstock et al.,2015; Steele et al.,2017) is useful but expensive
- High resolution satellite imagery are cheap but highly unstructured and contains measurement error.
  - Deep learning and neural network model often require a much bigger training set.

Framework

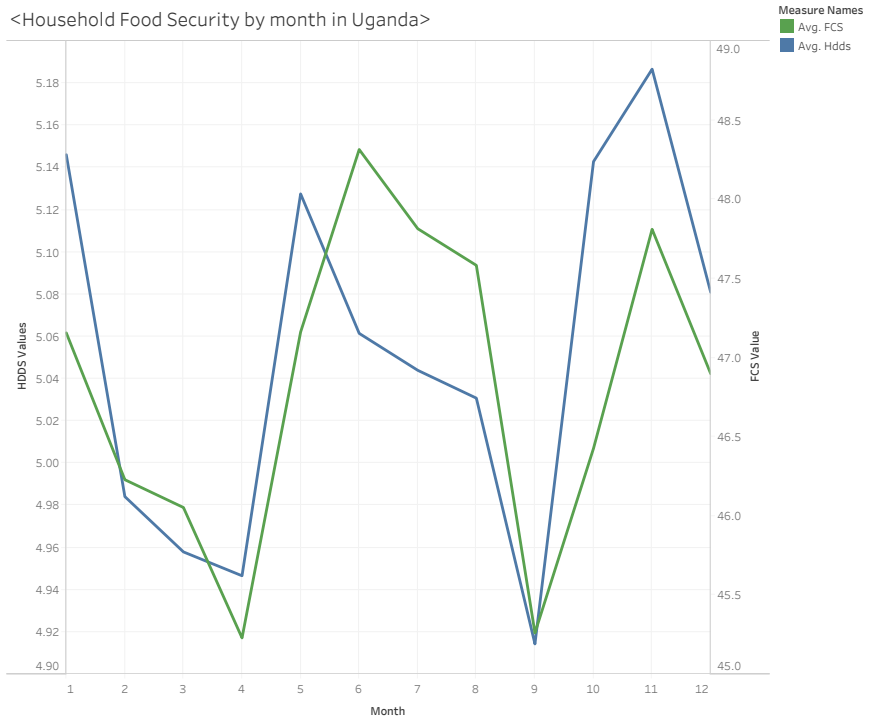
- Geo-referenced household surveys (LSMS) allow us to explore the spatial-temporal
- Align weather data with cropland calendar
- Align households to the most relevant price
-

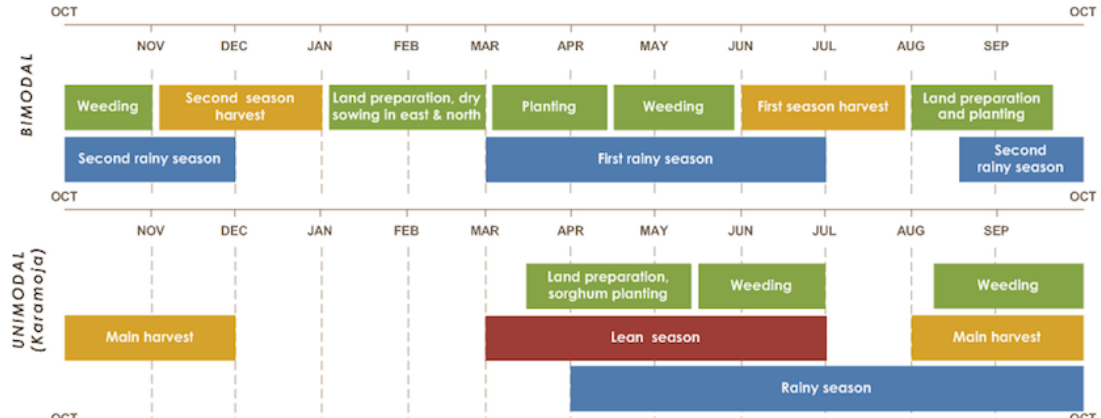
Temporal Variation





Temporal Variation





Spatio-temporal variation

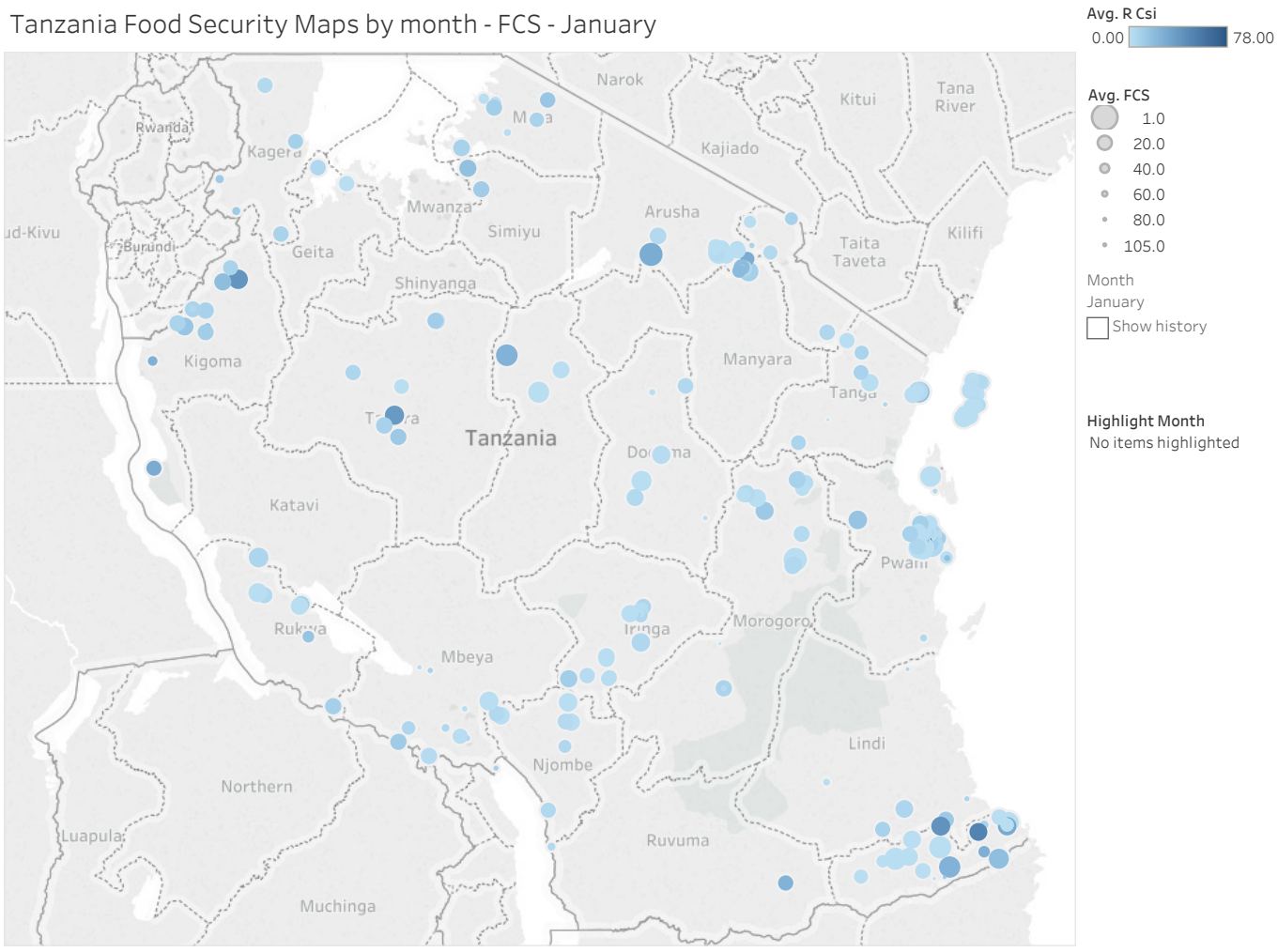


Figure 1: Food Security maps in Tanzania January

Spatialtemporal variation

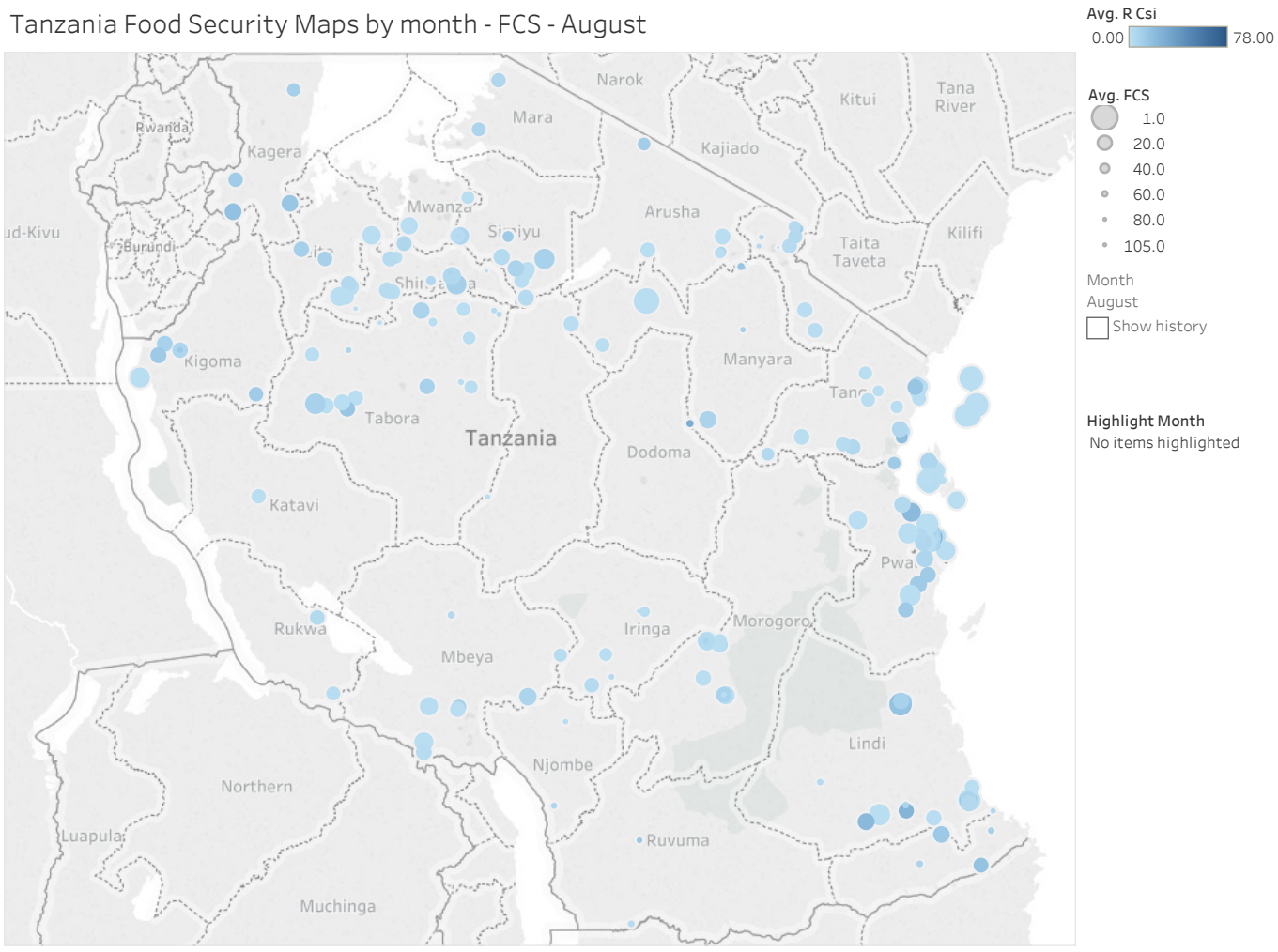


Figure 2: Food Security maps in Tanzania August

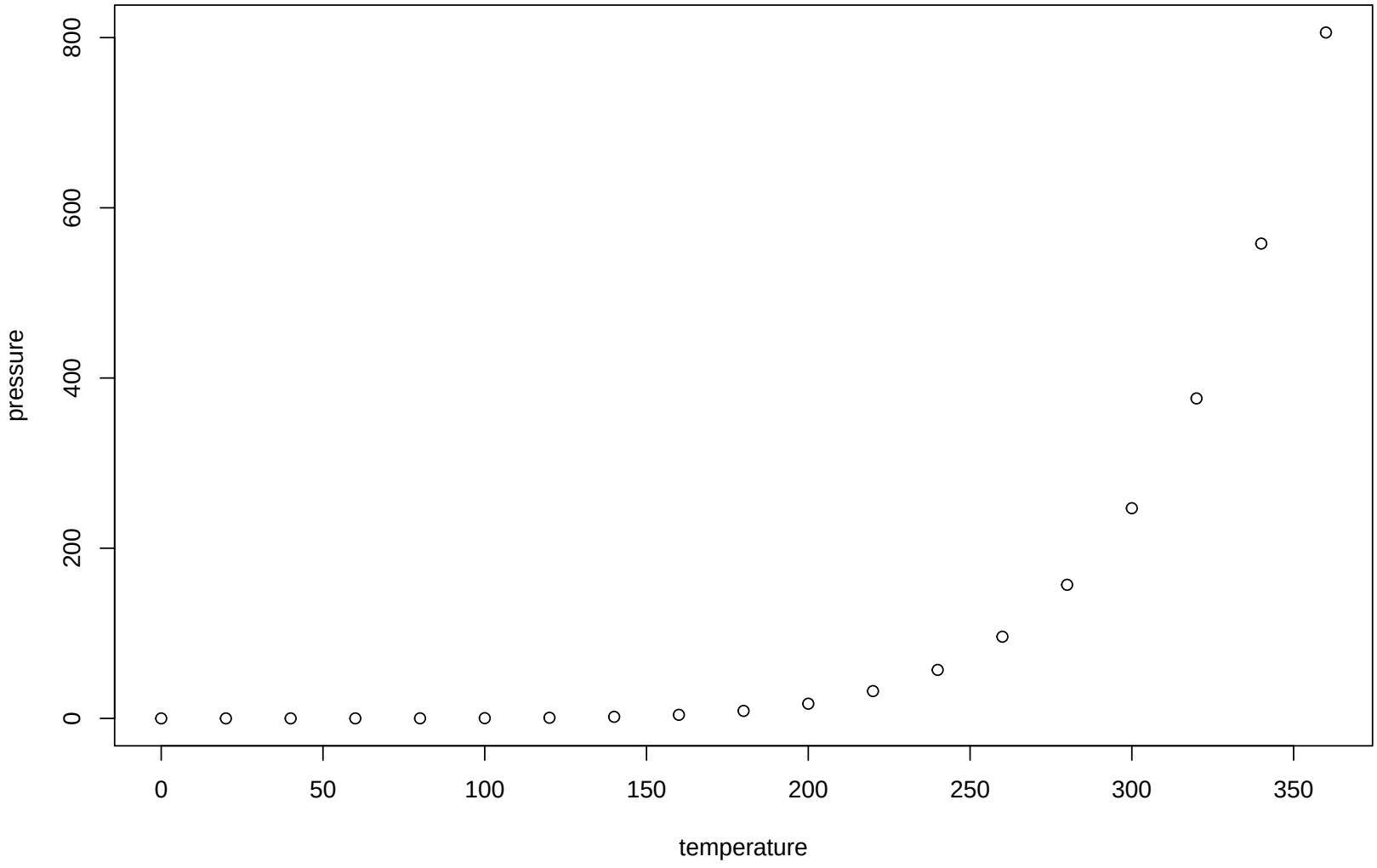
Discussion on the types of regression models

Data

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Modelling strategy



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)
-