Predicating Food Insecurity in Sub-Saharan Africa with Machine Learning

Application in Malawi and Tanzania

Yujun Zhou, Kathy Baylis

April 19, 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 to fill the data gap?
  - Price data of the main agricultural markets are collected monthly or weekly
  - Precipitation/temperature/soil quality from remote sensing are relevant to agricultural production
  - Distance to roads and markets reflects access to market and information
- Use supervised learning approach that to achieve higher predicative power and remain interpretable.

Preview of Results

- Out-of-sample predictions from our model explains up to 50%-67% of cluster level variations in Malawi and up to 76% in Tanzania for HDDS and FCS measures.
  - 0.4 r^2 on consumption expenditures and 0.6 r^2 on asset index (Jean *et al.*, 2016)
  - 0.76 r^2 on WI and 0.27 r^2 on income in Bangladesh (Steele JE *et al*,2017)
- Using the same sets of features, a machine learning model outperforms a baseline linear model by 20 - 50% in high dimensional settings.
- Decreased the proportion estimates that overestimate Food Security outcomes
- Validates the "A Prototype for Predicting Food Insecurity Using Readily Available Data" paper with Tanzania.

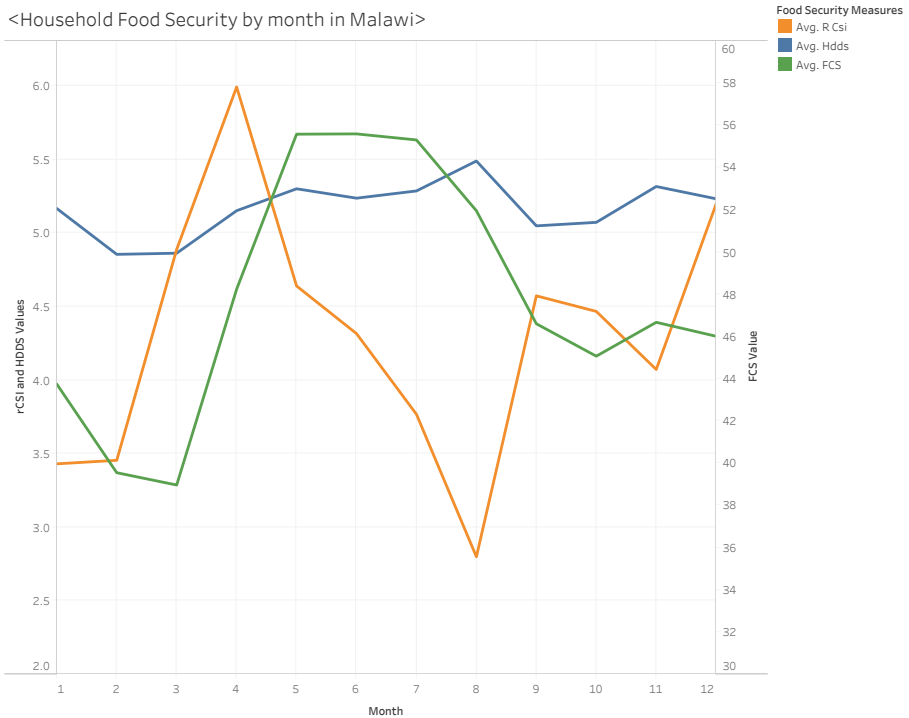
Literature Review

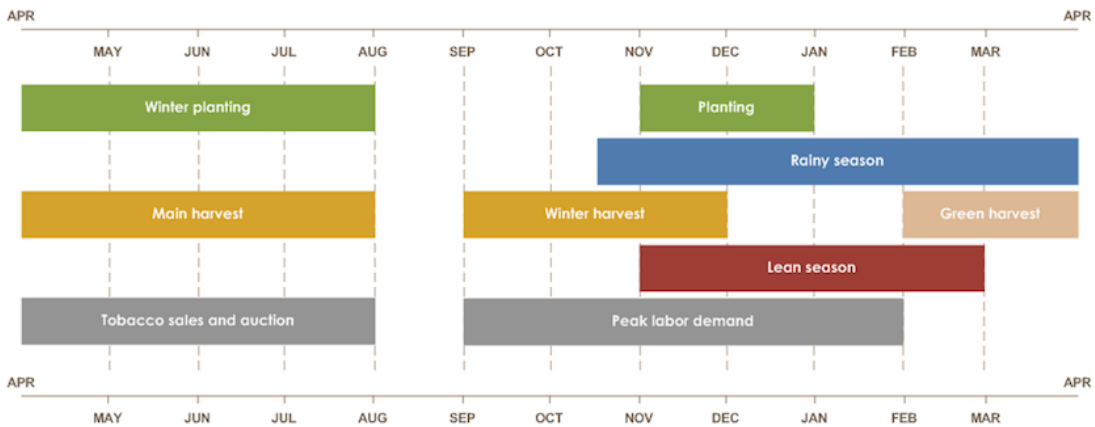
- Night lights data (Chen and Nordhaus 2011;Henderson *et al.* 2012.) does a good job at predicating economic activity but the variation is little in areas of the extreme poor or in urban areas.
- Mobile phone data (Blumenstock *et al.*,2015; Steele *et al.*,2017) is complimentary to remote sensing data but expensive.
- Very High resolution satellite imagery are becoming cheaper but highly unstructured and contains measurement error (Engstrom, 2018; 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.

Framework

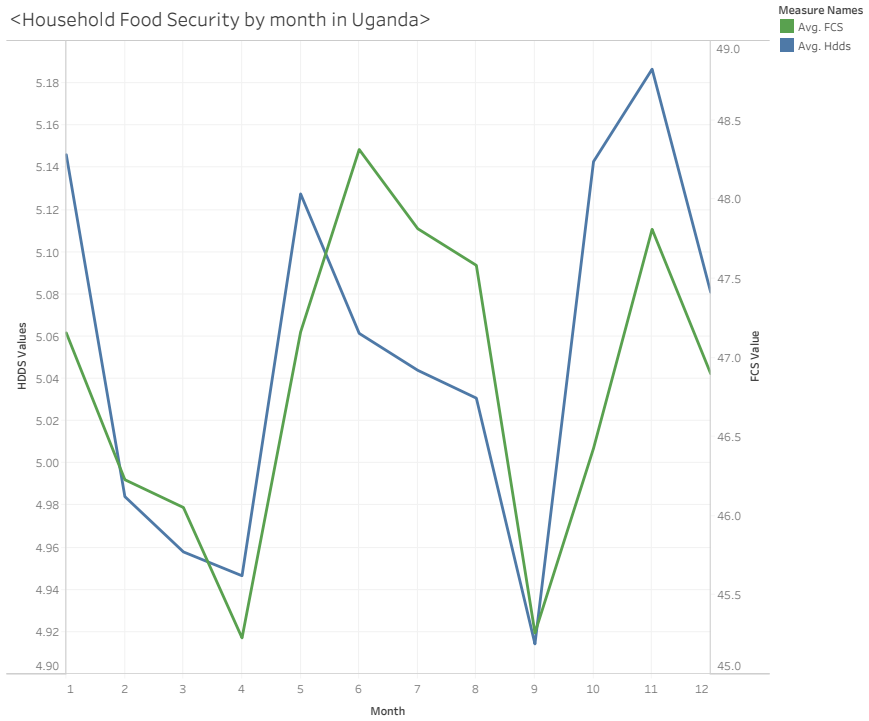
- Understanding of the food security. The definition of food security has various characteristics. The use metrics can lead to quite divergent rankings of the same population (Steele et al. 2017)
- Geo-referenced household surveys (LSMS data) allow us to explore the spatial-temporal variations in food 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
- Utilize the interaction and higher order terms of the variables

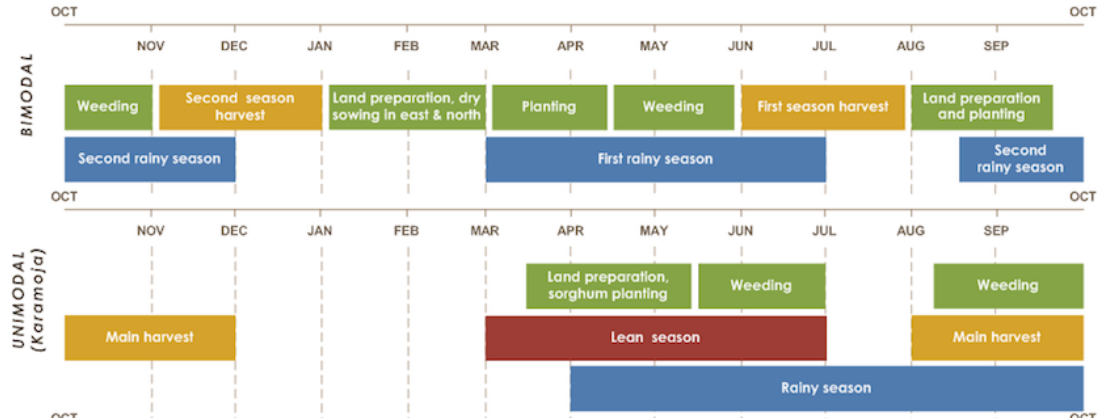
Temporal Variation





Temporal Variation





Spatio-temporal variation

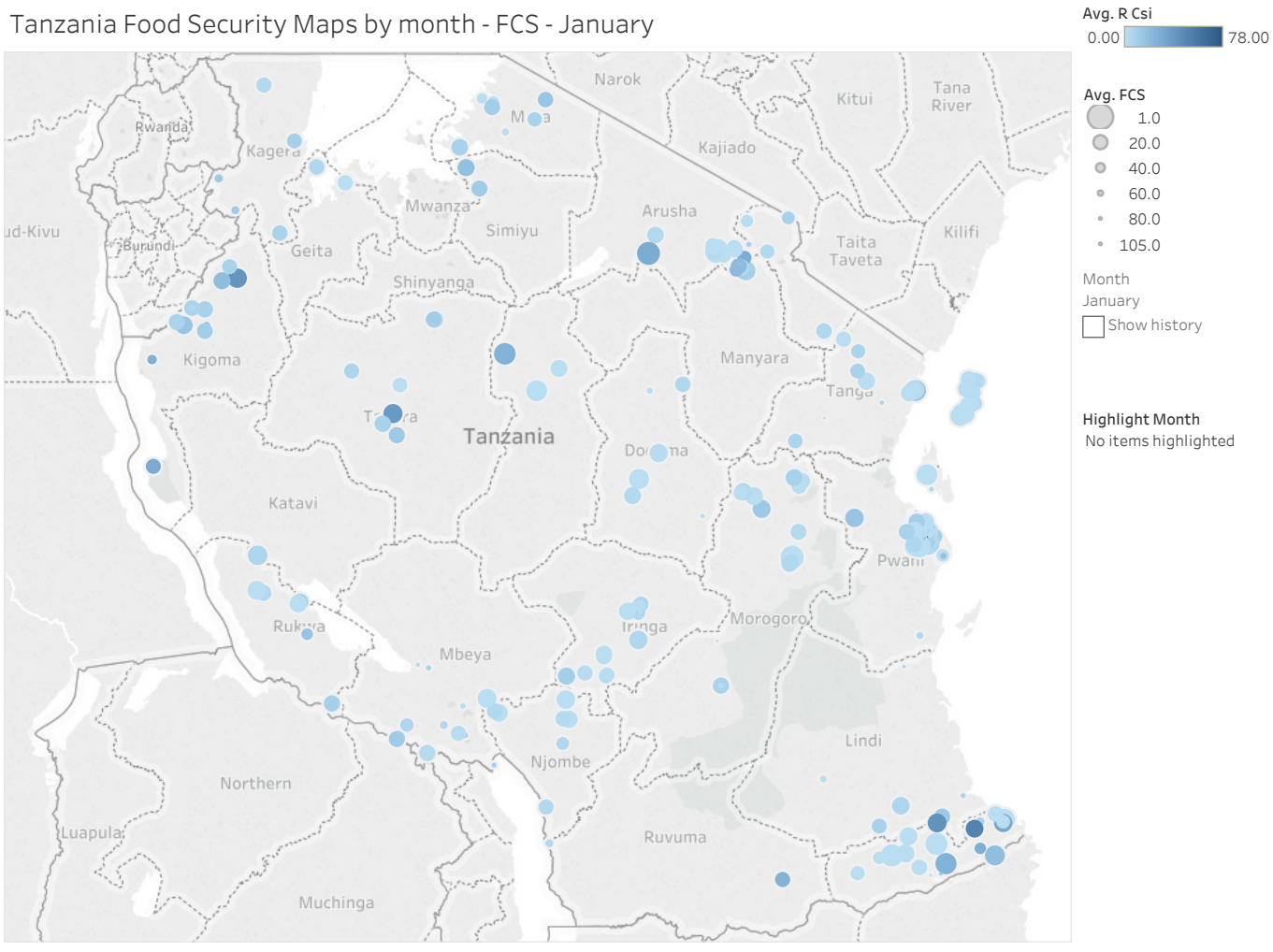


Figure 1: Food Security Maps in Tanzania, January

Spatialtemporal variation

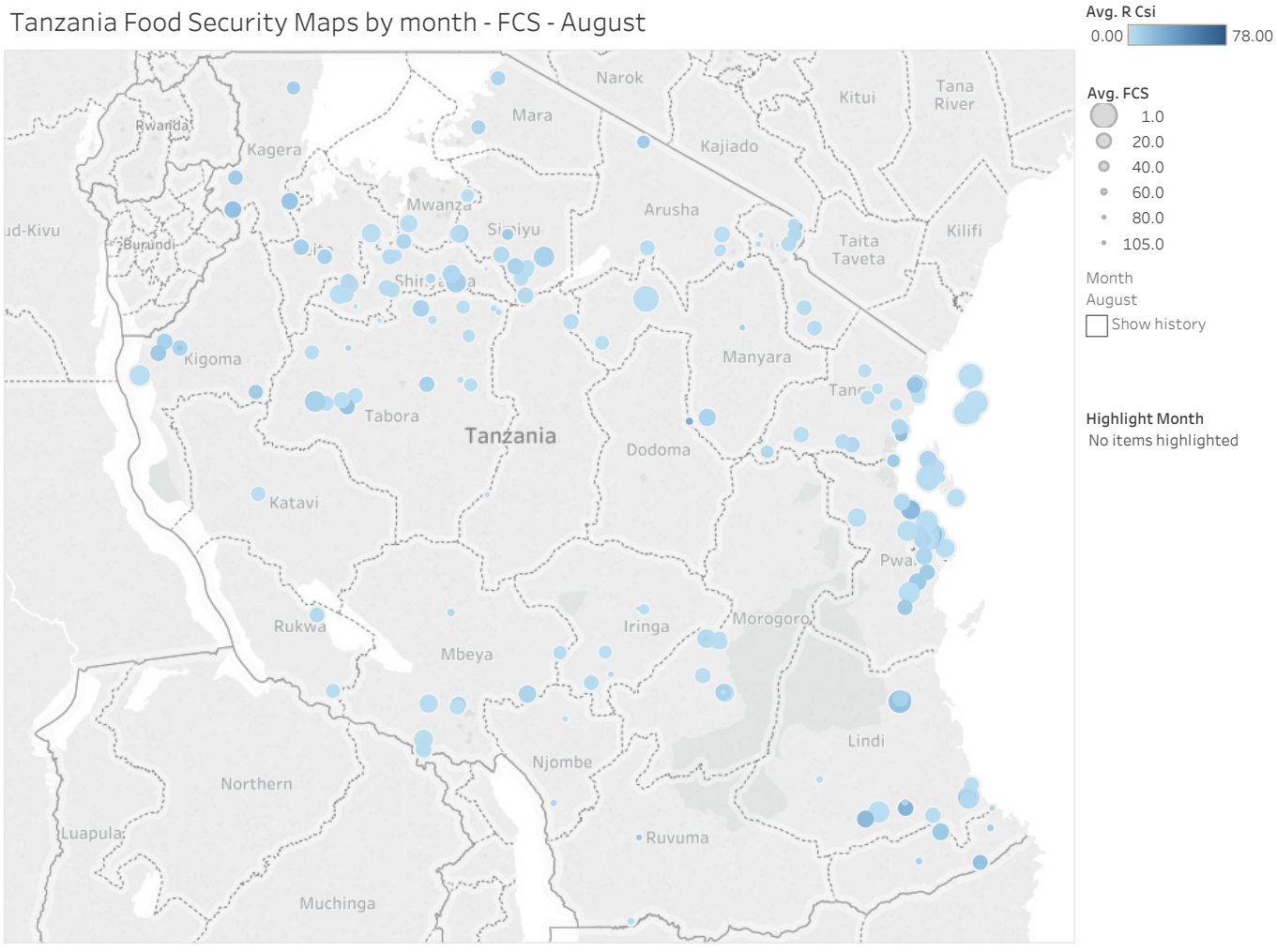


Figure 2: Food Security Maps in Tanzania, August

Data

- Price data (maize and rice)
- Weather
  - first day of rainfall
  - rainfall in growing season
  - maximum of days without rain
  - mean temperature
  - growing degree days (days that average temperature between 5-32 Degree Celsius)
- Soil quality: slope, water, soil nutrition
- Roof/floor type, household asset index
- Mobile phone ownership: access to financial resources, remittance flow and economic (Eagle *et al* 2010, Blumenstock *et al*)

Modelling strategy

- Set of models deal with high dimensional problem with different types regularization:
  - Lasso (L1 norm, prefers smaller model)
  - Ridge (L2 norm, prefers large model)
  - Elastic Net (Combination of the two)
- Set of models that are great with positively skewed and high dimensional data:
  - Random Forest
  - Gradient boosting

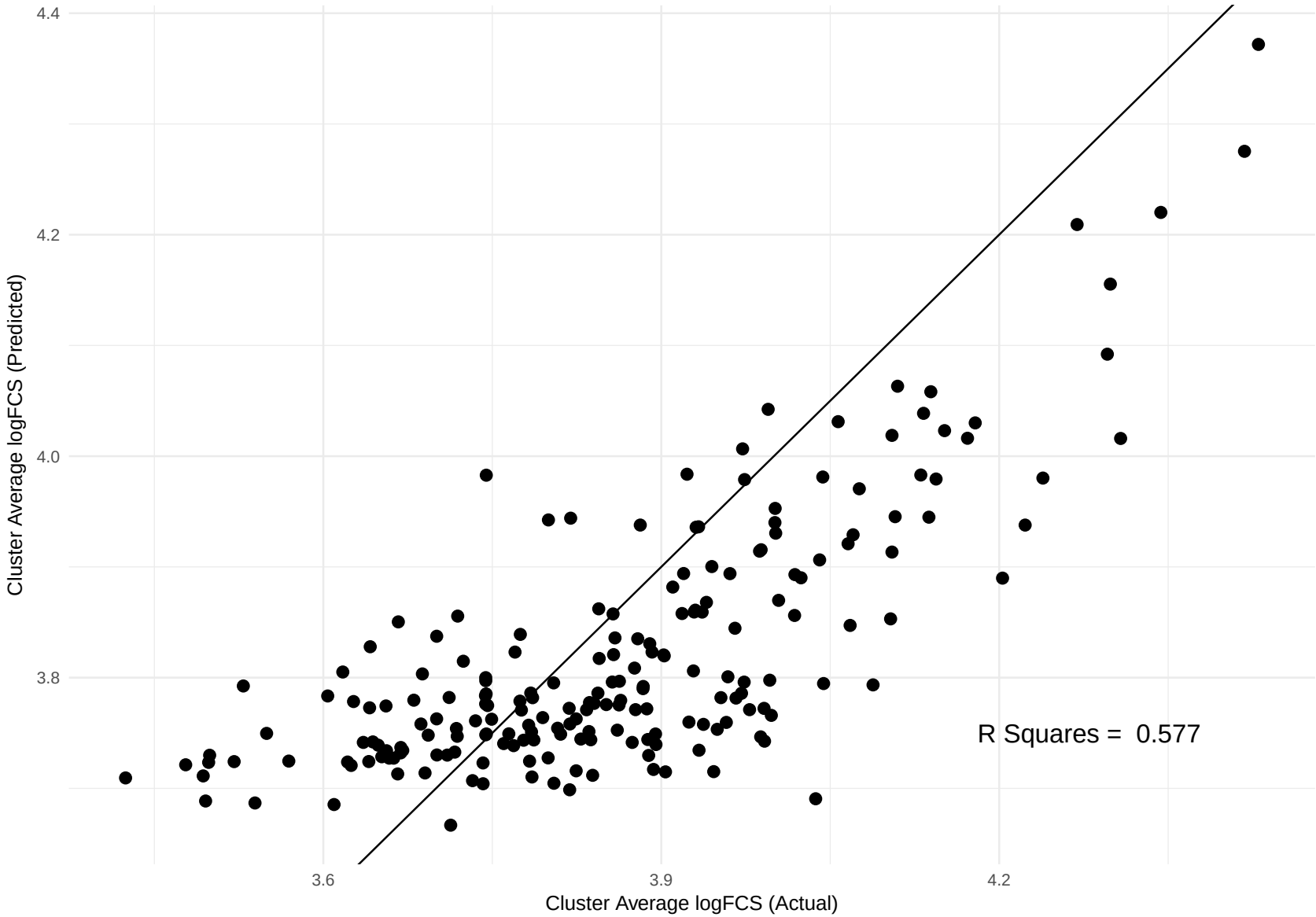
Main Results

Table 1: Cluster Level prediction R squares

Model	logFCS	HDDS	RCSI
Linear Model	0.5319	0.6730	0.0903
Linear Model with interaction terms	0.3140	0.2916	0.0530
Ridge	0.4650	0.6236	0.1230
BaynesianRidge	0.4660	0.6240	0.1230
Lasso	0.5770	0.6860	0.1420
ElasticNet	0.5760	0.6830	0.1200
GradientBoost	0.5767	0.6640	0.0660
Random Forest	0.5387	0.6470	0.0418

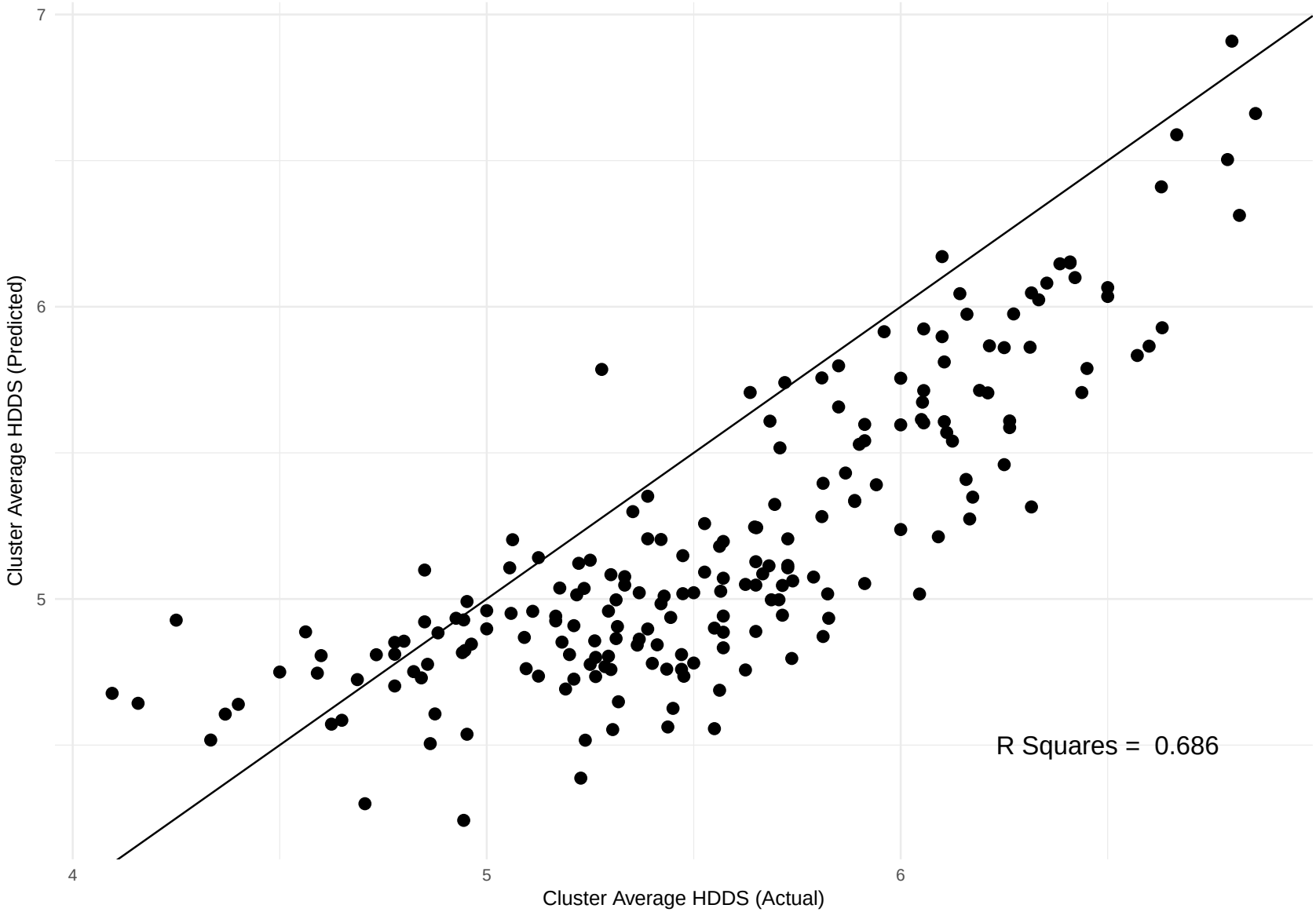
Scatter Plots

- FCS



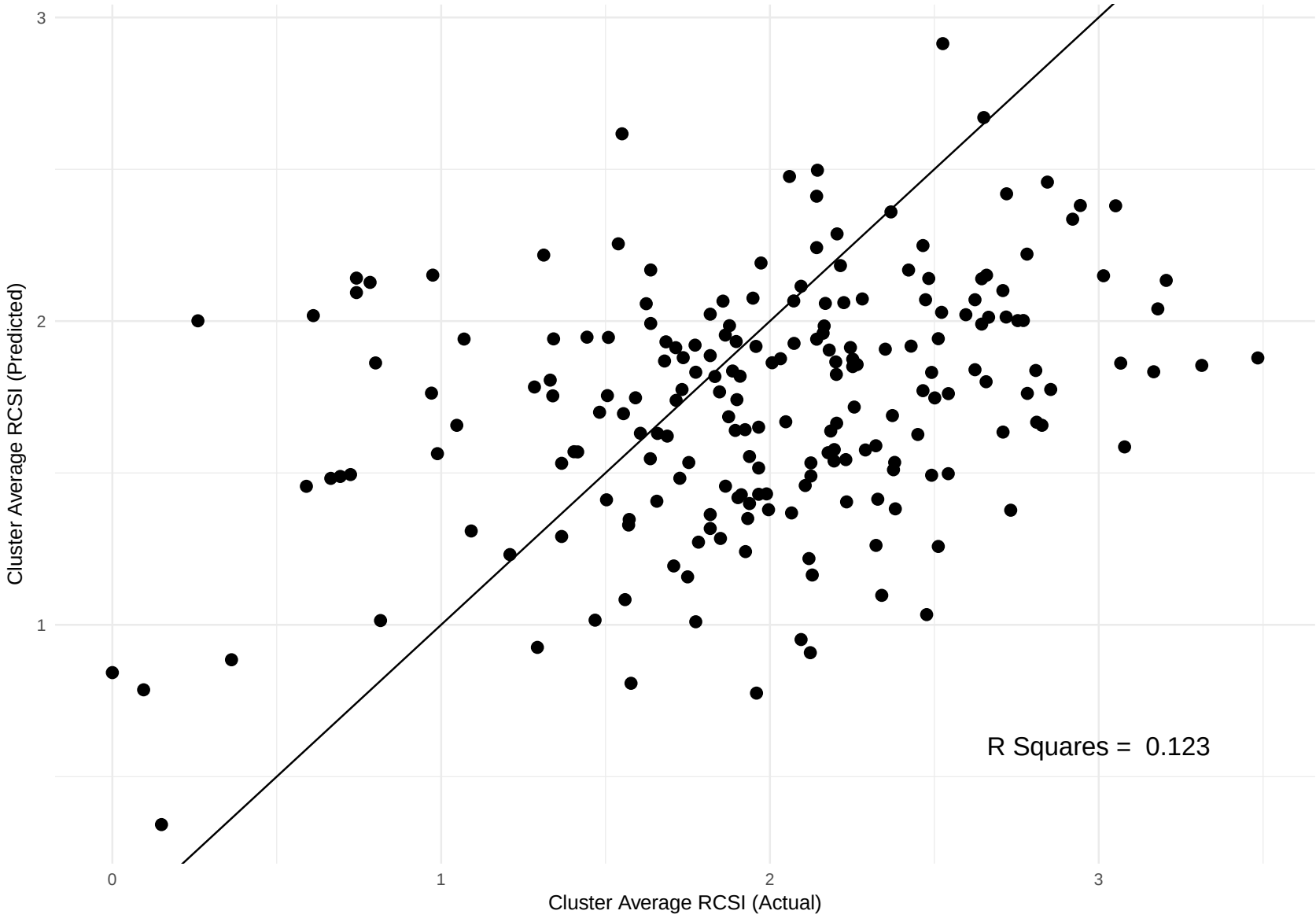
Scatter plots

- HDDS



Scatter plots

- RCSI



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

- limited to gradual food insecure cases and can't predict sudden, abrupt threat to food security (natural disaster, war and conflict)