**Predict Food Security with Machine Learning: Application in Eastern Africa**

*Submission to the Agricultural and Applied Economics Association Annual Meeting*

*May 2019 draft – please do not cite without permission*

**Yujun Zhou and Kathy Baylis**

***Selected Paper prepared for presentation at the 2019 Agricultural & Applied Economics Association Annual Meeting, Atlanta, GA, July 21-23***

**Introduction**

Faster response during food crises saves lives and resources. Crises are increasing in frequency and severity in many parts of the world. Identifying the scale and scope of these crises in a timely and accurate fashion is essential for food aid and humanitarian responses. However, policymakers often lack the information required to identify the right populations to target programming and resources (Barrett and Headey 2014). By 2012, only 27 of Africa's 48 countries had conducted at least two comparable household level surveys (Beegle et al. 2016) because it is costly to do so. The data gap hinders the efforts to effectively targeting the population in need and calls for the use of data and method that are cost-effective and accurate.

Novel data and data methods can be used to fill this data gap. Nightlights data (Chen and Nordhaus 2011; Henderson et al. 2012.) can serve as a proxy for economic activity, especially when comparing across countries. However, in remote rural or better off urban areas, the nightlight intensity varies little over time, hiding substantial changes in economic outcomes. Mobile phone data (Blumenstock et al.,2015; Steele et al.,2017) is more frequent and less expensive compared to census surveys. However, in the short term, it is not feasible to roll out cellphone surveys in entire sub-Saharan Africa, and the biases associated with using relying on cell phone-sourced information to infer population statistics are as of yet, not well understood. Very high-resolution satellite imagery is becoming cheaper but suffers the lack of structure (Engstrom et al., 2017; Donaldson and Storeygard, 2016). Recent studies have combined Convolutional Neural Network (CNN) models and transfer learning (Jean et al., 2016; Babenko et al. 2017) to make an inference based on the information in the satellite imageries. These models can explain up to 60% - 75% of the variation at the village level wealth and asset measures in several sub-Saharan Africa countries. However, the reliance on the information in the satellite imagery (specifically, building size, roof type, road conditions) limits its performance on development indicators other than wealth or assets. Head et al. (2017) apply the Jean et al. (2016) approach to a set of various development indicators and across several countries. Their research finds that the prediction performance degrades quickly on health and nutrition outcomes (no better than random guessing in some cases). The reliance on nightlight data on this approach also limits the prediction accuracy when applied in countries with different socioeconomic conditions. The external validity and interpretability of this deep learning-based approach call for a method tailored for food security predictions.

This paper uses data-driven framework to predict the onset of food crises. Combining remote sensing data with household surveys and price data, the model is able produces the most spatially and temporally granular predictions of food security. With an emphasis on the structure of the prediction error, this paper uses various machine learning techniques to increase the accuracy and reducing the type II error in predicting food security status. The empirical application of the

method is in Malawi, Uganda and Tanzania, using the Living Standard Monitoring Survey (LSMS) as the reference data.

**Data:**

We plan to use readily available data to model the food security status of village clusters in Uganda, Malawi, and Tanzania. We predict three measures of food security used by international humanitarian organizations including USAID and the World Food Programme (WFP): the reduced coping strategies index (rCSI), the household dietary diversity score (HDDS) and the food consumption score (FCS).

The variables used to predict food security are high-frequency data including precipitation, temperature, market prices, soil quality and geographic variables, which are generally collected remotely and are widely available. Household roof type is used as a crude proxy of poverty that can be accurately captured from satellite imagery. Cellular phones are access to financial resources, market information and remittance flow (Eagle et al. 2010, Blumenstock et al. 2016) also serve as important predictors. Household-level data including demographics and assets from LSMS are also included.

**Method**

The models are trained using a training set in one year. The accuracy of the models is evaluated using out-of-sample data in another year.

Our model tries to explain these variations in food security by the spatial-temporal variation in food availability and food access. Specifically, we align weather data with the crop growing season to describe the temporary shocks in food availability. We also align households with their most relevant market price, as shocks to income and household consumption budget.

To deal with the high dimensionality and nonlinearity problem, we apply machine learning methods including regularization methods (LASSO, Elastic net) and Ensemble learning models (Random Forest and Gradient boosting) to improve prediction accuracy. Another issue with the prediction accuracy is that the model works well on the quality measures of food security (FCS and HDDS), but a lot worse on the quantitative measure (rCSI). Partially due to the survey questions to the construction of the variable, the distribution of the rCSI appears to be highly skewed, a long-tail with a mass of points around zero. We employed oversampling and down sampling techniques to improve prediction performance.

The choice of the best model is the key in this paper. The criterion for choosing the best model is a fixture of prediction accuracy, recall rate (reducing type II error) and model interpretability.

## Initial Results

Currently, the prediction accuracy in around an r squared of 0.6– 0.7 at the cluster level and around 0.4 at the household level. Preliminary results suggest assets related variables are explaining most of the variance in food security: cellphone ownership, asset index, and roof types. Market access variables, for example, distance to road and distance to markets are also important. Total rainfall, temperature and the start of the rainy season play a big role as well.

## References


Babenko, Boris, et al. "Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico." *arXiv preprint arXiv:1711.06323* (2017).

Barrett, Christopher B., and Derek Headey. "A proposal for measuring resilience in a risky world." (2014).

Beegle, Kathleen, et al. *Poverty in a rising Africa*. The World Bank, 2016.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. "Predicting poverty and wealth from mobile phone metadata." *Science* 350.6264 (2015): 1073-1076.

Castelluccio, Marco, et al. "Land use classification in remote sensing images by convolutional neural networks." *arXiv preprint arXiv:1508.00092* (2015).

Chen, Derek. "Temporal Poverty Prediction using Satellite Imagery." (2017)

Chen, Xi, and William D. Nordhaus. "Using luminosity data as a proxy for economic statistics." *Proceedings of the National Academy of Sciences* 108.21 (2011): 8589-8594.

Dang, Hai-Anh, Dean Jolliffe, and Calogero Carletto. "Data gaps, data incomparability, and data imputation: a review of poverty measurement methods for data-scarce environments." (2017).

Donaldson, Dave, and Adam Storeygard. "The view from above: Applications of satellite data in economics." *Journal of Economic Perspectives* 30.4 (2016): 171-98.

Engstrom, Ryan, Jonathan Hersh, and David Newhouse. "Poverty from space: using high-resolution satellite imagery for estimating economic well-being." (2017).

Head, Andrew, et al. "Can Human Development be Measured with Satellite Imagery?." *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. ACM, 2017.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. "Measuring economic growth from outer space." *American economic review* 102.2 (2012): 994-1028.

Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." *Science* 353.6301 (2016): 790-794.

Kussul, Nataliia, et al. "Deep learning classification of land cover and crop types using remote sensing data." *IEEE Geoscience and Remote Sensing Letters* 14.5 (2017): 778-782.

Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences, 114*(46), E9783-E9792.

Steele, Jessica E., et al. "Mapping poverty using mobile phone and satellite data." *Journal of The Royal Society Interface*14.127 (2017): 20160690.