**Predict Food Security with Machine Learning Proposal**

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 builds on the data-driven framework in Lentz et al. (2018) 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 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 Malawi and Tanzania. We briefly describe our food security measures and the three classes of data below.

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). These measures capture different dimensions of food insecurity; rCSI measures household coping strategies associated with available food quantity, while FCS and HDDS reflect dietary quality. The HDDS is a count of the number of food categories that a household consumes in a week. The FCS weights this count of food groups according to their nutrient density. Higher values of both measures indicate higher food security. Because the rCSI measures the number of coping strategies a household uses to address possible food shortages, a higher rCSI indicates lower food security. Most often, governments and international agencies apply cut-offs to categorize food security status rather than use the continuous measures of food security.

These three classes of variables increase in processing requirements and a decrease in availability.

Class 1 data are high-frequency data including precipitation, market prices, soil quality and geographic variables. Class 1 data are generally collected remotely and are widely available. Agricultural prices are often readily available and collected in person by government agents although can also be collected using cellular phone technologies. We generate measures of agriculturally-relevant precipitation generated from the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) dataset. We use the total amount of precipitation that fell during the October–April rainy season in Malawi and Tanzania. For the same season, we define the length of the most extended dry spell as the number of continuous days with no rain. To measure the beginning of the rainy season, we calculate the number of days after October 1st in which rainfall greater than 10 mm fell three days out of five. These three variables are taken from the prior agricultural season to predict food availability for the June/July maize harvest. We also include the maximum amount of rainfall in the current month to control for possible flooding, which can affect transportation and local economic outcomes. For the temperature aspect, we include mean temperature and growing degree days (days that average temperature between 5-32 Degree Celsius) proposed by Deschenes and Greenstone (2007).

Class 2 data are data likely to be available but also likely to require additional work to be accessed and processed, such as data on household roof type (a coarse predictor of poverty) and cell phone ownership. Mobile phone ownership: access to financial resources, remittance flow and economic (Eagle et al. 2010, Blumenstock et al. 2016). For the Class 2 data, we draw from the LSMS. We include the percent of households who own a cell phone and the roof type (metal versus thatch), both of which could be collected from sources other than recent surveys, such as cellular companies and through remote sensing.

Class 3 data consist of infrequently gathered but publicly available household-level data including demographics and assets. These data are often available in a census or large household surveys such as the Demographic Health Surveys or LSMS. Data for our Class 3 variables also come from the LSMS surveys: demographic data, including the gender and age of the household head, number of household members, and assets.

Method:

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

Geo-referenced household surveys (LSMS data) allow us to explore the spatial-temporal variations in food security measures. We can observe a nationally representative sample in different months and across agroecological zones. Our model tries to explain these variations in food security by the spatial-temporal variation in food availability and food access. Precisely, we align weather data with the crop growing season to explain the temporary shocks in food availability. We also align households with their most relevant market price, as shocks to income and household consumption budget. The feature selection process also includes the interaction and higher order terms of all variables from class 1 to class 3.

Many regressors in the model may overfit the training set and leads to lowered accuracy in out-of-sample predictions. To deal with the high dimensional problem, we apply machine learning methods with regularization constraint on the maximizing problem in estimation, such as Lasso, Ridge and Elastic Net.

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 measure, the distribution of the rCSI appears to be long-tailed with a mass of points around zero. Fitting a linear regression on a variable with this long-tailed distribution would be predicted correctly in the mean of the distribution yet tend to over-prediction on the tail side of the distribution. The type II error in classifying food security status, in this case, can be a concern in targeting the right population that needs aid in times of a food crisis. A possible solution is to use a mixture model by splitting the distribution and fit models separately. Methods related to generalized Pareto distribution may be another solution. Ensemble learning models (Random Forest and Gradient boosting) are also great in dealing with positively skewed and high dimensional data.

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

Preliminary results suggest assets related variables are explaining most of the variance in food security: cellphone ownership, asset index and roof types. In sparse models, only the assets variables are left in the models. Market access variables, for example, distance to road and distance to ADMARC (Agricultural Development and Marketing Corporation) markets are also important. Total rainfall, temperature and the start of the rainy season also matters.

Currently the best performing models is Gradient boosting and Gaussian Process, reduces prediction error for the safer households significantly. Having problem reducing the error on predicting more endangered households. Lasso and Random Forest perform similarly but are more interpretable with variables of importance listed.

Next steps:

The possible next steps can come from the following aspects.

1. Add a nonlinear term f (x) in the equation and use a Gaussian Process (a nonparametric approach) to estimate it. Choose a proper kernel for the Gaussian Process. Pokhriyal & Jacques (2017) incorporates spatial coordinates into the kernel to reflect spatial autocorrelation.
2. (worth trying) Use a mixture model (i.e., estimating the function as the sum of two Gaussians) for different types of data. In the case of rCSI, maybe the sum of a Gaussian and a log-normal distribution to help deal with the long-tailed distribution.
3. Compare the results of classification algorithm after binning the target variables and regression variables with continuous variables. Discussion should put emphasis on improving the recall rate of the food insecure category.
4. Use a validated set to select features in different countries to allow for different variables across country.

**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.