**Predict Food Security with Machine Learning**

**Outline**

1. **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 a 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 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.

The current food security early warning system uses the Integrated Food Security Phase Classification System (IPC). The lack of near-term, sub-national predictions delays effective response. Lentz et al. (2019) present a data-driven approach to predict the onset of food crises that combines remote sensing data with household surveys and price data. The model is able produces the most spatially and temporally granular predictions of food security and achieves high prediction accuracy overall.

Following the framework in Lentz et al. (2019), this paper incorporates data on food security, rainfall, and prices to predict the onset of food crises. We expand the study area to Malawi, Uganda, and Tanzania to test the framework with more heterogeneity in geography, environment, and socioeconomic status. With an emphasis on capturing the majority of food insecure households, this paper uses various machine learning techniques to increase the performance in detecting the insecure food clusters. We aim to further develop the data flow into a system that is automatically updated, generalizable, scalable, and cost-effective in predicting areas of potential food shortage.

1. **Data**
2. Food security measurements

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).

1. Explanatory variables:

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. 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. We also align households with their most relevant market price, like shocks to income and household consumption budget.

Household-level data, including demographics and assets from LSMS, are also included. Household roof and floor type are 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 significant predictors.

1. **Method**
2. Categorical vs. continuous: focus on the categorical prediction for the given cutoffs of each food security measures
3. Result metrics: recall, precision, AUROC, instead of accuracy
4. Cost-sensitive learning: penalize misclassifications of the minority class more heavily by having a cost function, which is equal to the inverse of the class proportions
5. Sampling: down-sampling, over-sampling, and SMOTE
6. Classification Algorithms:

a. Tree-based: Classification Tree (baseline and base learner), Random Forest, Extreme Gradient boosting

b. Anomaly detection: clustering methods, One-class SVMs, and Isolation Forests

6. Data split: Year split, regional split, random split

7. Data segmentation: by country, one dataset, by rural/urban, auto-segmentation

1. **Results**

1. Table 1: Baseline vs. ML algorithms, on several result metrics

2. Table 2: Baseline vs. ML algorithms with sampling technique

3. Parameter tuning and feature importance (appendix)

4. Table 3: Data split comparisons

5. Table 4: Data Segmentation Comparisons

6. Table 5: Error analysis: type of clusters, regions, month

7. Figure 1: Food security map of the three countries

8. Figure 2: Feature importance related to tree splits

9. Figure 3: Optimal data segmentation

**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.