Thesis Proposal

Forecasting Poverty with Machine Learning

Dang Ngoc Huy
h.dang@mpp.hertie-school.org

Advisor: Slava Jankin

Hertie School of Governance
Master of Public Policy Analysis

1 Literature Review:[1]

In 2015, the United Nations adopted 17 Sustainable Development Goals, one of which is the complete eradication of poverty by 2030. To realize such an ambition, the first step is essentially to identify the areas most affected by poverty, to better target aid and development programs. However, the business of mapping poverty is a costly one with complex politics, limited and unreliable availability of information. To resolve this identification problem, Marshall Burke and his team of researchers at Stanford's University's Department of Earth System Science proposed the novel approach of analysing satellite images of the planet's surface to pinpoint regions that might be regarded as ravaged by poverty in their seminal work: **Combining satellite imagery and machine learning to predict poverty**, Science, vol. 353, no. 6301, pp. 790–794, 2016.

To train a model that can identify impoverished areas, Burke and his teams utilized three different data types: images of luminosity at night time which could signal economic activity level; images of daytime from which useful information on socioeconomic data could be extracted; and finally survey data on household consumption expenditure and asset wealth as validation for the findings extrapolated from the analysis of satellite images.

The learning process occurred in two consecutive steps: the satellite images of daytime and night-time are first utilized to create segmentations of land use, signs of human activities and other useful geographic information; in the second phase, the survey data is fed into the convolutional neural network model which then can learn the patterns and connection between the different features that it had identified from the images and the household spending and asset possession. Focusing on five countries from Africa - Nigeria, Tanzania, Uganda, Malawi, and Rwanda, the researchers were able to explain up to 75% of the variation in economic outcomes at the local level just based on the satellite images which were freely available in the public domain.

With a model that now has an understanding of the correlation between the features present in the satellite images and poverty level, the research could then map out predictions on regions where there are gaps in the survey data, helping to identify previously unknown areas for aid targeting.

2 Motivation:

Measuring poverty to better target aid and development is a difficult business for a number of reasons: the reliance on the collection of household data which is time-consuming and costly; household surveys and data are often not available for many countries of interest and when they are, it is usually not at the desired frequency and this time-lag creates problems in the decision making process for development and aid agencies; to get a sense of a region's current development, it is often essential to use predictions to fill in the gap in the time-series data of that region's economic indicators. Building on the research by Burke et al., the key questions of my research will, therefore, be:

How Machine Learning and Deep Learning can help to identify and understand areas of poverty? There are currently 3 approaches that I want to research further on:

- How to utilize classification and clustering algorithm to identify and categorize poverty;

[1] Project GitHub: https://github.com/huydang90/Forecast-Poverty-With-Machine-Learning

- How to utilize regression algorithm to forecast consumptions and expenditures which are often used as the indicators of poverty. Being able to classify a household as poor or non-poor is but one piece of the puzzle as most development programs are quite nuanced in their definition and has a range of spectrum on what is considered as poverty;
- How to utilize Deep learning with Convolutional Neural Networks and Transfer Learning to analyze satellite images and identify or estimate areas of poverty through identification of land use, light pollution and other predictors;
- How to combine all the approaches above into a new metric to compliment the current methods of mapping and forecasting poverty?
- What features identified by the model from the satellite images are the best predictor of poverty level? Many different features and geographic information can be extracted from the images such as road, buildings, vegetation, land usage which can contribute to the accurate forecast of poverty level. Evaluating how each of these features can impact the model prediction could contribute further to the understanding of poverty and its identification strategy.

3 Task:

To achieve the goals of the research, the following tasks should be implemented:

- Review and learn about state-of-the-art in image processing, convolutional neural network for image segmentation, geospatial information system;
- Replicate baseline model from Burke et al.
- Identify main country of research
- Collect data: satellite images and economic indicators data of the interested country;
- Construct probabilistic graphical model and model architecture of CNN;
- Experiment with modeling;
- Write up the thesis with the results obtained from modeling and experimentation;

4 Data:

I intend to use the following data resources for my research:

- World Bank, African Development Bank, Asian Development Bank data on poverty levels and their related predictors in different countries. I will need to select certain countries with different levels of class balance and imbalance as examples to test run the algorithms and models. All banks have data portals and libraries that can be accessed by the public for research. However, whether these resources will contain the necessary data for this particular research is yet to be determined;
- Administrative data on economic indicators from the targeted country of research;
- Satellite images data of regions of interest which can be accessed in open-sourced GIS projects such as Google Satellite Image, USGS Earth Explorer, Copernicus Open Access Hub (access to satellites Sentinel-1, -2 and -3), and NASA EarthData Search.

5 Method:

As aforementioned, an important part of my dissertation will be to seek an understanding whether Machine Learning and Deep Learning can be viable instruments for the toolbox of forecasting and measuring poverty. To this end, the following methodology will be utilized:

- Machine Learning: using traditional algorithm for regression tasks to identify and forecast poverty based on only a handful of easily collected variables;

- Deep Learning: using Convolution Neural Network and Transfer Learning to extract information and find patterns in satellite images to predict poverty level;

6 Baseline:

The baseline of the research will be the multistep learning approach by Burke et al.

7 Successful outcome:

The application of these technologies in this development area is relatively new and it is debatable regarding how effective, accurate and scalable these technologies can be in comparison to current methodologies currently in use for poverty analysis, such as SWIFT (Survey of Well-being via Instant and Frequent Tracking), especially in data-scarce environments. Or perhaps they are all legitimate techniques that can be utilized in specialized use cases.

Therefore, a good outcome would be successful implementation of machine learning approach for satellite image processing with reasonable accuracy in a completely new country that have not been used for research, to understand how generalizable this methodology can be and whether governments can reliably utilize it to supplement their anti-poverty efforts.

8 Reference:

N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, *Combining satellite imagery and machine learning to predict poverty*, Science, vol. 353, no. 6301, pp. 790–794, 2016.

9 Timeline: