Poverty Estimation Using Satellite Images and Geospatial Data

Christine Cepelak
216776@mds.hertie-school.org

Janine De Vera
janinepdevera@gmail.com

Johannes Halkenhäußer
halkenjo@gmail.com

1. Introduction

1.1. Background

Poverty is a multifaceted problem manifested by broad conditions such as malnutrition, homelessness, lack of access to clean water, and low educational achievement. It continues to be one the world's most pressing issues and it has only been exacerbated by the economic effects of the COVID-19 pandemic. The United Nations' Sustainable Development Agenda for 2015-2030 identifies poverty eradication as one of its top priorities – evidence that governments across the globe are committed to alleviating poverty.

It follows that poverty statistics are among the most important and most widely used data in the economic and policy research sphere. Practical and ground-level applications of poverty data include identification of vulnerable populations for rolling out interventions by humanitarian and development organizations. However, measuring and monitoring poverty is both conceptually and operationally challenging. Poverty measurement is commonly the responsibility of national statistics offices, but the capacity to produce accurate and frequent information varies widely depending on the country. Ironically, the most vulnerable economies in need of extensive and up-to-date poverty data are also those who lack the capacity to compile them. Countries in extreme poverty or in conflict often lack relevant survey data for several years.

This project[1] aims to contribute to the research body on bridging poverty statistics measurement gaps, through the application of machine learning techniques in satellite images and open source spatial data. The proposed machine learning models will incorporate intuitive geospatial variables to facilitate insightful interpretations and comparisons of different factors associated with poverty.

1.2. Methodology Overview

To measure actual on-the-ground situations, data from official Demographic and Health Surveys (DHS)[2] or population censuses will be collected. As in related studies [2, 3, 4, 5], a wealth index or poverty index will be constructed as the variable of interest and primary measure of poverty and economic wellbeing. This will serve as the ground-level truth that will be used in conjunction with geospatial covariates such as nighttime luminosity, population mobility, accessibility, connectivity, and consumer preferences. Information will be collected and processed based on pre-defined geographic clusters or grids within multiple countries of interest.

As a starting point for research, seven countries from the most poverty-stricken regions in the world have been selected:

- Guatemala (South America)
- Haiti (Carribean)
- Kenya (Africa)
- Mongolia (Central Asia)
- Nepal (South Asia)
- Pakistan (Middle East)
- Timor-Leste (Southeast Asia)

The goal is to find complete and representative data for each country, so the focus of the study may shift to neighboring economies as the project progresses.

Three models will be estimated using data from the chosen countries:

1. *Baseline model*: A simple linear regression examines the effect of identified explanatory variables on the wealth index. [3, 4] A slightly more complex variation is a spatial regression that accounts for effects of geographical characteristics of observations.

2. *Random forest*: A random forest regression estimates multiple regression trees using different combinations of variables and averages over several predictions. The primary objective of such an approach is improving model explainability and robustness. [3, 4].

3. *Deep learning algorithm*: A Convolutional Neural Network (CNN) is commonly used for machine vision

[1] GitHub respository: https://github.com/ccepelak/ML-SS22

[2] USAID DHS database: https://www.dhsprogram.com/Countries

algorithms. Unlike regular neural networks CNNs use a convolutional and a pooling layer to extract information from a subspace of the image thus preserving some of the underlying structure of the image. [5]

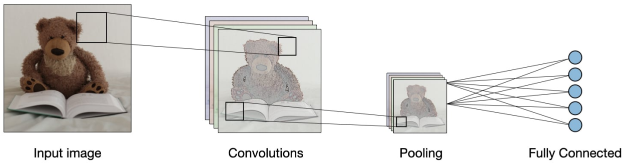

Figure 1. The essential architecture of a convolutional neural network [1].

2. Motivation

As the world wrestles with the impacts of AI advancements and hyper connectivity on social media platforms, a basic foundation of governance is missing in many countries: accurate census data. As stated in the Background, the critical nature of addressing poverty needs no explanation; however, the urgent need to apply modern technological advances to basic social-political administration receives notably less attention.

We hope these models can be applied to some of the most impoverished countries in the world. For example, Somalia and Afghanistan are in the small minority of countries which have not had a national census since 1990. Neighboring countries such as Pakistan and Kenya, however, have more substantial data available, which can be used to evaluate predictions.

Without this type of census and poverty data, not only do countries lack the means to meaningfully support their own communities, but the international community cannot lend aid.

3. Evaluation

Our motivation informs the goal outcome: a model with strong out of sample prediction and high external validity. An ideal model could hence be used to give an indication of where poverty is prevalent in countries that have outdated or poor quality census and survey data.

Success will mean to train a model that is in the realm of explanatory power as [2, 3, 4] which achieve a R^2 of 0.66 and 0.70 respectively. The literature repeatedly refers to the R^2 as the evaluation metric, which we will hence adapt to evaluate our models likewise. The R^2 measures the variance in the dependent variable that can be explained by the independent variables. We will also use our model and predict poverty scores in a country not included in the training data and again report the R^2.

4. Resources

There is a wealth of geospatial information that could potentially lend the proposed models more predictive power. Examples of geospatial variables and their data sources are summarized in Table 1. The list is preliminary and not exhaustive.

Variable	Measurement	Sources
Economic activity	Nighttime lights satellite images from remote sensing data	Google Earth Engine, QuickBird Imageries
Population mobility	Population movement across different categories of places (e.g. workplace, residential, groceries and pharmacies, transit stations)	Google Mobility Report
Accessibility	Distance from cluster centers to primary and secondary roads; counts of points of interests such as banks, schools, hospitals	OpenStreetMap
Connectivity	Internet performance, stability, and speed	Ookla Open Datasets
Consumer characteristics	Consumer goods preferences; number of Facebook users with breakdown per access segment (i.e. 4G, 3G, 2G, WiFi)	Facebook API

Table 1. Geospatial variables and data sources.

We will rely heavily on common machine learning libraries such as pandas and scikit-learn (data cleaning/wrangling, preprocessing, feature engineering, linear model, random forest) and keras (CNN).

Granular satellite image data is quite large and hence we will preferably use the Google Data storage provided to us. Computationally, the final models should be trained using Hertie's GPUs. Setting up the pipeline can be done in Google Colab or on our devices with a subset of data.

5. Contributions

We expect to work in close collaboration on realising our project. To that end we have set up a standing weekly call and will be splitting our workload according to our strengths. All three of us will work on procuring and harmonizing our data. To ensure that our models stay comparable, the pre-processing will also be worked on collaboratively while Janine and Christine will work on creating a clean data set, Johannes will work on the visualization of the data as well as relaying potential feature engineering or data anomalies back to the group. The models will be split by previous experience. Christine and Janine will develop the linear regression and the Random Forest model. Johannes has previously worked on NN and will focus on developing the convolutional neural network (CNN). The write-up/presentation will again be worked on together, yet each member of the group will be writing about the model they developed so no knowledge is lost between coding and write up.

To ensure timely completion of the project, all members will follow the proposed timeline presented in Table 2.

Date	Deliverables
Preparatory work	
wk 1 Mar	Proposal submission; data gathering; finalization of countries to be included in study
wk 2 Mar	Data gathering and cleaning
Exploratory analysis	
wk 3 Mar*	Exploratory data analysis; write up of initial findings
wk 4 Mar	Finalization of mid-term report; Model 1 (Linear Regression)
Modeling	
wk 1 Apr	Model 2 (Random Forest), Model 3 (CNN)
wk 2 Apr*	Model 2 (Random Forest), Model 3 (CNN)
Report and documentation	
wk 3 Apr	Finalization of analysis; documentation on Github; write up of final report
wk 4 Apr	Submission of final report; preparation for final presentation
wk 1 May	Presentation

* Weeks when consultation meetings will be scheduled

*

Table 2. Proposed timeline of deliverables.

References

[1] S. Amidi and A. Amidi. Convolutional neural networks cheatsheet, 2019.

[2] X. Chen and W. D. Nordhaus. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594, 2011.

[3] C. Ledesma, O. L. Garonita, L. J. Flores, I. Tingzon, and D. Dalisay. Interpretable poverty mapping using social media data, satellite images, and geospatial information. *arXiv preprint arXiv:2011.13563*, 2020.

[4] X. Zhao, B. Yu, Y. Liu, Z. Chen, Q. Li, C. Wang, and J. Wu. Estimation of poverty using random forest regression with multi-source data: A case study in bangladesh. *Remote Sensing*, 11(4):375, 2019.

[5] Z. Zhongming, L. Linong, Y. Xiaona, Z. Wangqiang, L. Wei, et al. Mapping poverty through data integration and artificial intelligence: A special supplement of the key indicators for asia and the pacific. 2020.