Takaaki Masaki (The World Bank), David Newhouse (The World Bank)

Using Big Data for Small Areas: An Application in Tanzania

Understanding poverty at the local level has become increasingly important given a sign of widening income inequalities within countries and increasing need for better targeting projects or investments to reach the poorest. Obtaining accurate and reliable estimates of local poverty, however, is difficult due to the high costs of collecting welfare data that allows for such analysis. Household surveys – from which poverty estimates are derived – are typically too small to produce reliable estimates below a certain geographical level, such as provinces or districts.

To address this challenge, this paper applies a variant of the Elbers, Lanjouw, and Lanjouw (2003) method (Silwal et al, forthcoming). The traditional application of the ELL method estimates a model that relates welfare to both household and community characteristics in the welfare survey (Elberts, Lanjouw, and Leite, 2008). This model is then applied to the estimated parameters to the target census data to simulate the distribution of welfare in the full census. This traditional approach, however, imposes the strong assumption that household characteristics used as predictors in the first-stage consumption model in the survey and in the census, such as household size and education, are drawn from the same distribution. This assumption is often untenable if there is a substantial temporal gap between the source and survey data, or if the questions are asked in significantly different ways, which could bias the poverty estimates. Additionally, the traditional ELL limits the set of household variables that can be used for consumption/poverty estimation to those present in both the survey and census data.

This paper applies the ELL framework but – instead of relying on the census data – we harness satellite-based geospatial data. One of the key advantages of satellite or remote-sensing data is that they are often available globally and capture important variation in the geographical characteristics of small areas or communities that are proven to be correlated with welfare. The proposed approach employs spatial data at the village level both to predict consumption in the modeling stage and to simulate the distribution of poverty for the entire population, which by construction meets the assumption of traditional ELL that the predictors in the source and target data are drawn from the same distribution. While using spatial data at the village instead of household data reduces the precision of the estimates, it still provides a meaningful gain over the use of survey data alone.

We apply our proposed version of ELL to produce estimates for 169 Tanzanian districts by combining data from the 2018 household budget survey (HBS) and a battery of spatial data derived from the Sentinel 2 satellite and other publicly available geospatial indicators. The most recent estimate of local (or district-level) poverty in Tanzania employs the traditional application of the ELL and relies on the HBS 2011/12 and the short questionnaire of the Population and Housing Census (PHC 2012). Despite its usefulness in guiding policies to reduce poverty, the 2012 poverty map has become obsolete and may no longer reflect robust economic growth that Tanzania has experienced over the past seven years. The application of our proposed estimation strategy is particularly relevant to Tanzania where the new HBS has just been conducted (in 2018) and the six year temporal gap between this survey and the last census (2012) is to raise questions about whether the assumptions of traditional ELL are met.

Our application of ELL leverages satellite or remote-sensing data – instead of household-level characteristics – to construct a consumption model and then use it to simulate the distribution of welfare across all households. In the first stage of the consumption model, the satellite data are summarized to the lowest level of aggregation that is available to be matched accurately with the geographical locations of households in the HBS – which, in this case, means 16,351 villages. Geospatial data on the right hand side of equation are drawn from a number of different sources – including, but not limited to, nighttime light data from the Visible Infrared Imaging Radiometer Suite (VIIRS), built-up area data from Global Human Settlement Layer (GHSL), population data from WorldPop, as well as spatial features extracted from a cloud free mosaic of 2017/2018 Sentinel-2 imagery.

This paper shows how our proposed ELL application with the spatial covariates (without relying on the census data) improves the accuracy and reliability of district-level poverty estimates substantially compared with survey-based estimates alone. Spatial data can offer a strong complement to the census data and significantly improve the precision of local poverty estimates. In this paper, we also test the performance of our proposed approach by conducting ten-fold cross-validation and also testing its precision/accuracy against other commonly utilized methodologies in small area poverty estimation (e.g., Fay-Herriot (1979), Battese-Harter-Fuller (1988), and Hierarchical Bayes). The paper concludes by highlighting the comparative advantages and disadvantages of our proposed method vis-à-vis others and identifying key areas that call for further research.