2019
IARIW-World Bank

Special IARIW-World Bank Conference "New Approaches to Defining and Measuring Poverty in a Growing World" Washington, DC, November 7-8, 2019

The Spatial Distribution of Poverty in Sri Lanka in 2016

Ani Rudra Silwal
David Newhouse
Ryan Engstrom
Adane Bedada
Dilhanie Deepawansa

Paper Prepared for the IARIW-World Bank Conference
Washington, DC, November 7-8, 2019
Session 2B: Data Methods for Improved Poverty Measurement
Time: 14:15 – 16:15, September 7

The spatial distribution of poverty in Sri Lanka in 2016[1]

Ani Rudra Silwal, David Newhouse, Ryan Engstrom, Adane Bedada, Dilhanie Deepawansa[2]

PRELIMINARY DRAFT. PLEASE DO NOT CITE.
April 2019

[1] We would like to thank Indu Bandara, Paul Corral, Roy van der Weide, Keith Garrett, Sashikala Jeyaraj, Yeon Soo Kim, Nandini Krishnan, Tatiana Nenova, Minh Cong Ngyuen, Idah Pswarayi-Riddihough, Benjamin Stewart, and Thomas Walker for their support at various stages of this work. We thank Thomas Walker for his help securing funding for this project, which was provided by the World Bank's Rapid Social Response Multi-Donor Trust Fund.
[2] Silwal and Newhouse are with the World Bank's Poverty and Equity Global Practice. Engstrom and Bedada are at the George Washington University. The corresponding author is David Newhouse(dnewhouse@worldbank.org).

1. Introduction

Reliable poverty statistics derived from household surveys are typically only available for large administrative areas. Samples are typically too small to generate reliable estimates at lower administrative levels secause surveys are costly. However, policymakers and aid agencies increasingly demand timely and granular poverty statistics to identify pockets of poverty and allocate scarce public resources within a country. Small area estimation techniques borrow strength from auxiliary data with greater coverage to generate more precise estimates of poverty for small geographic areas or demographic groups. Traditionally, this exercise relies on estimating a model in a household survey that predicts welfare based on household characteristics and applies the estimated coefficients and parameters to census data to simulate poverty rates. A key requirement of this method is that the household characteristics in the census and the survey are drawn from the same underlying probability distribution. This is a strong assumption, especially if key questions are measured differently, or if there is a long gap between the census and the survey.

In Sri Lanka, the first poverty map conducted in 2002 was a powerful tool in measuring and comparing poverty at disaggregate administrative levels (World Bank 2005). An important application of this map was to inform policy makers during the reform of the Samurdhi transfer program in 2005, when the Ministry of Samurdhi used the map to identify the poorest Divisional Secretariat Divisions in the country. The widespread acceptance and use of the second poverty map for 2012/13 is a testament to the success of Sri Lanka's Department of Census and Statistics (DCS) in disseminating the results of the poverty mapping exercise throughout the government agencies as well as to the general public.

Despite its usefulness in guiding policies to reduce poverty, the 2012/13 poverty map has become outdated and may no longer reflect recent developments in Sri Lanka's economic conditions. The end of the civil conflict in 2009 has led to heavy public investment in the ex-conflict areas in the North and East, as well as the opening of that area to trade with India. The rest of the country has also seen heavy investments in infrastructure and social protection. The national economy has also grown at a modest rate, although not all parts of the country may have benefited equally. A new poverty map, therefore, can inform policy makers whether previous pockets of poverty have persisted and whether new pockets have emerged. An update of Sri Lanka's poverty map based on the newly available 2016 HIES is particularly timely because the new map will shed light on poverty trends in the Northern and Eastern provinces since 2012/13, which were not included in the 2002 map due to the lack of available census data.

The goal of this exercise is to generate small area estimates of poverty in Sri Lanka in 2016 that incorporates the latest available data and methodological advances.[3] The major challenge is that the census and the survey were conducted four years apart, making it unrealistic to assume that household characteristics in the survey and the census are drawn from the same statistical distribution. Therefore, it is not appropriate to directly apply the methodology proposed in ELL (2003) and used in the 2002 and 2012/13 poverty maps.[4]

[3] For example, it relies on the new *sae* package developed for Stata, rather than the standalone Povmap2 software used in the past (Nguyen et al. 2018).

[4] We are not aware of any hard and fast rule on how small the temporal gap between the survey and census must be for traditional ELL to be valid. However, the World Bank recommends comparing the means of common variables in the survey and census and not proceeding with traditional ELL if the means in the census tend to fall outside the confidence interval of the survey means.

We address the temporal gap between the census and the survey by excluding household characteristics from the consumption module. We instead use geographic-means of variables from the census, following Lange, Pape, and Wollberg (2018), as well as remote-sensing data. Because these indicators are linked with the household survey geographically, they can be used both to develop the prediction model and simulate welfare in the census. In other words, we build a model that predicts household consumption, taken from the HIES 2016 data, as a function of geographic means of census and remote census variables. The model colefficients and parameters are then used to predict consumption for all census households. The variables used in the model stage are therefore the same variables, drawn from the same distribution, as those used in the simulation stage. One consequence of this method is that the explanatory power of the consumption model is lower than a model that incorporates household characteristics, because the explanatory variables can only capture variation across villages. As a result, subdistrict estimates obtained when excluding household predictors are less precise, and it is crucial to carefully evaluate the precision and reliability of the estimates using this method.

We combine data from a national census, a nationally-representative survey, and a variety of features derived from satellite imagery, linked at the village level, to generate sub-district poverty estimates for Sri Lanka in 2016. Incorporating census and remote sensing indicators reduces the average coefficient of variation of subdistrict poverty estimates by approximately one third compared with direct survey-based estimates, which is roughly equivalent to doubling the effective size of the sample. The predictions derived from satellite data generate much more plausible estimates of changes in subdistrict poverty rates than the traditional method of using household characteristics in the model. The results strongly suggest that indicators derived from satellite imagery or other geographically linked data can be combined with survey data to better track changes in poverty in local areas, even in the absence of a new census.

The rest of this note is organized as follows. Section 2 describes the census, survey, and remote sensing datasets we use to construct a poverty map for Sri Lanka in 2016. Section 3 describes the empirical methodology. Section 4 presents the main results, robustness checks, and a description of changes in the spatial distribution of poverty since 2012. Section 5 summarizes the main findings and lessons.

2. Data

The main challenge in generating reliable poverty statistics for small administrative areas is that welfare data is costly to collect, and therefore samples tend to be too small to generate statistically reliable estimates. Small area estimation surmounts this challenge by "borrowing strength" from data sources with broader coverage, such as census data. Linking the census data with the survey, through a regression model, generates estimates of poverty in areas covered by the census but not the survey. Incorporating these estimates in areas not covered by the survey, in turn, makes small area estimates more precise and reliable. In this case, the goal is to generate reliable estimates of poverty at the subdistrict level, which in Sri Lanka are called Divisional Secretariat Divisions (DSDs). There are 331 DSDs which are further divided into 14,022 Grama Niladhari Divisions (GNDs).[5] We use three sources of data for this analysis: The 2016 Household Income and Expenditure Survey 2016, the 2012 Census data, and remote sensing features derived from satellite imagery. The remainder of this section describes these data

a. Household Income and Expenditure Survey (HIES) 2016

[5] The average population of a GND and DSD is about 1,450 and 62,000, respectively.

The HIES is a nationally representative survey conducted by the Department of Census and Statistics of Sri Lanka approximately every three years for repeated cross-sections of the Sri Lankan population. The HIES is a multi-topic survey that also contains extensive modules on household consumption and income. This survey is used to estimate the official poverty statistics in Sri Lanka and was last conducted in 2016 with a sample size of 21,756 households.[6] The sample size was selected such that the coefficient of variation (CV) of estimated average household consumption for each district would be approximately 10 percent, which is the lowest level that the survey is considered to be representative. The HIES 2016 sample includes households in 329 DSDs and 2,491 GNDs and thus covers about 18 percent of the GNDs in the country. We drop one outlier DSDs whose average per capita consumption is 3 standard deviations below the mean per capita consumption of all 329 DSDs. This outlier DSD is Island North (Kayts) DSD in Jaffna District which contains 16 sample households 75% of whom have per capita consumption below the national poverty line.[7]

b. Census of Population and Housing 2012

The last Sri Lankan census was conducted in 2012 and covered the entire country.[8] The 2012 Census contains the usual questions on the demographics of all members in the household, assets owned by the household, and characteristics of the household dwelling. The final Census dataset used for this exercise includes 5.23 million households with 19.74 million individuals. The official population estimate in 2012 was 20.36 million, thus our analysis only covers 96.97% of the official population count. The primary reason for this discrepancy is that identifiers were occasionally missing in the individual, household, and housing units. In generating the household database for this exercise, we drop all individuals that are not usual residents, all collective living quarters, and non-housing units. Some of these categories may have been included in the official population count.

c. Remote sensing data

Satellite imagery is an exciting new source of data for the estimation of geographically-disaggregated poverty statistics on a more frequent basis. The availability of satellite imagery and the ability to process it has improved significantly in recent years. Satellite imagery has extensive spatial and temporal coverage, especially for recent years, and captures a significant portion of spatial differences in living standards across villages in Sri Lanka (Engstrom et al, 2017) and Africa (Jean et al, 2016). We use two sets of remote sensing variables in the analysis: Features derived from publicly available data products, and spatial features calculated from a processed version of publicly available Sentinel-2 imagery. In each case, features were computed at both the GN and DS levels:

The five features derived from publicly available remote data products were:

- **Rainfall** data from Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) collected by NASA (Funk et al, 2015). We generate the mean, standard deviation, and Z-scores (normalized deviation from historical mean for 1981-2015) for each quarter of 2016.

---

[6] The national poverty line of Sri Lanka is Rs. 4,166 per person per month, which translates to $2.48 per day in 2011 PPP terms. The national poverty line was last set in 2002 using the prevailing consumption basket; it has since been raised by considering changes in the consumer price index.

[7] It is possible that the primary sampling unit selected for HIES 2016 in this DSD happens to be unusually poor. The survey-based estimated poverty rate for this DSD is 78% at the national line. This is extremely high, compared with the the 7.7% poverty rate for Jaffna District in 2016 and the 15.3% estimated poverty rate for this DSD in 2012/13.

[8] March 20, 2012 was designated as the official census day. (Source: Census of Population and Housing 2012: Provisional Information Based on 5% Sample).

- **Night-time lights** for 2016 from the Visible Infrared Imaging Radiometer Suite (VIIRS) at a spatial resolution of 15 arc-seconds.
- **Elevation and slope** data from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) sensor at a spatial resolution of 30 meters.
- **Global forest cover change** data based on Hansen et al. (2013), from which we use gain and loss in forest area between 2000 and 2014, at a spatial resolution of 30 meters.
- **Global human settlement layer** (GHSL): This dataset contains built-up areas on a global scale for the following years: 1975, 1990, 200, 2014. It is published by the Joint Research Center (JRC) of the European Commission and is derived from data collected by Landsat satellites. We compute the percentage of total built-up area observed in 2014 that was constructed prior to 1975 or during 1975-1990, 1975-1990, and 2000-2014.

Spatial features were derived from a cloud free mosaic of -2017-2018 Sentinel-2 imagery, which is collected every 5 days by Sentinel 2 sensors on board two satellites, Sentinel 2A and 2B. This imagery is made publicly available by the European Space Agency. Sentinel 2 imagery is the highest spatial resolution (10m per pixel) optical data that is publicly available. While the spatial resolution is generally not sharp enough to distinguish individual buildings, it can distinguish inhabited from uninhabited areas and the spatial pattern within the built environment. Sentinel 2 images from January 1, 2017 to March 31, 2018 were combined on a per-pixel basis in Google Earth Engine to create a cloud free mosaic of the entire country. The dates for imagery collection were chosen to account for significant cloud cover issues over tropical, Sri Lanka and the fact that Sentinel 2B was not launched until March 7, 2017. Spatial features were then calculated using Sp.Feas, an open-source Python library for processing contextual image features from satellite imagery.[9] These contextual features have been shown to be strongly correlated with poverty and population density (Engstrom et al, 2017, 2018). The contextual features used are as follows:

- **Fourier Transform** (FT) which is used to detect high or low frequency of lines.
- **Gabor Filter**, a linear Gaussian filter used for edge detection (Mehrota et al, 1992)
- **Histogram of Oriented Gradients** (HOG), which captures the orientation and magnitude of the shades of the image. (Dalal and Triggs, 2005)
- **Lacunarity** (LAC), which describes the extent of gaps and holes in a texture. Low-lacunarity geometric objects are homogeneous because all gap sizes are the same, whereas high-lacunarity objects are heterogeneous. (Myint et al, 2006).
- **Line Support Regions** (LSR), which characterize line attributes (Ünsalan and Boyer, 2004)
- **Normalized Difference Vegetation Index** (NDVI), the most widely used vegetation index that provides information about the health and amount of vegetation
- **PanTex**, which is a built-up presence index derived from the grey-level co-occurrence matrix (Pesaresi et al, 2008)
- **Structural Feature Sets** (SFS), which are statistical measures to extract the structural features of direction lines (Huang et al, 2007)

These spatial features are calculated by comparing pixels with their neighbors and then reporting this value back to the individual pixel (in this case 10m). The number of neighboring pixels considered in the comparison is the scale, which varies by the feature being calculated. In this analysis, we use scale of 3, 5, 7, and squares of 3 pixels by 3 pixels, 5 pixels by 5 pixels, to 7 pixels by 7 pixels for most features. This constitutes looking at an area of 30, 50, and 70 meters. For SFS, Fourier and LSR the scale was increased by a factor of 10 because these features need more area to capture the variation in the landscape. When applied to satellite imagery, the features capture "textures" of neighborhoods and are hence called

[9] Sp.feas is deveoped and maintained by Jordan Graesser. See https://github.com/jgrss/spfeas

contextual features. Once calculated the contextual features are then aggregated to construct GN division indicators, by taking the average and standard deviation within each GN Division. As an example, the Pantex feature captures the minimum contrast between a pixel and its neighbors. Highly built-up neighborhoods tend to have greater contrast in all directions, which will create high values of this feature. In contrast, in rural areas the pixel's brightness will likely be similar to a neighbor in at least one direction, which will create a low minimum contrast.

These features are primitive versions of the features constructed using machine learning techniques such as Convolutional Neural Networks (Jean, et al, 2016). Each approach summarizes images by comparing pixels with their neighbors. The main difference is that the Convolutional Neural Networks require survey data on welfare to determine which features to calculate. In other words, the computer selects parameters for layers of filters, which when applied to the imagery construct textures that are optimized to distinguish between low and higher welfare areas. For the computer to select the best parameters for these layers, millions of data points are typically used to train the algorithms. In contrast, the contextual features used in this analysis are constructed using pre-determined algorithms. Therefore, they are independent of the survey data and do not require confidential, geo-referenced survey data to train the algorithms. While these contextual features may not predict as well as convolutional neural networks, the relative performance of the two approaches when applied to household survey data containing tens of thousands of training observations is not currently known. Additionally, contextual features have the advantage of being much simpler and cheaper to apply than Convolutional Neural Networks.

3. Methodology

Direct estimates of poverty based on a household survey become less reliable when applied to lower administrative levels. As a result, national statistical offices only report statistics on poverty up to a certain administrative level for which the statistics are deemed to be sufficiently reliable. In the case of Sri Lanka, DCS only reports poverty statistics at the national and district levels. However, policymakers could benefit from knowledge of the variation in poverty across DSDs within district. The goal of this exercise is to improve the precision of DSD-estimates of poverty so that they are judged by the Department of Census and Statistics to be sufficiently reliable to be used for policy-making.

a. Linked ELL approach

Our methodology relies on a slight extension to the small area estimation method developed by Elbers, Lanjouw, and Lanjouw (2003). The ELL method generates the joint distribution of a welfare variable and independent variables in a dataset which does not have the welfare variable, such as a census, by estimating the relationship between these variables in an alternate dataset, such as a household survey. This method has been widely used by both the World Bank and international researchers to estimate poverty at disaggregated administrative areas in various countries and involves two stages.[10]

Traditionally, ELL estimates a model using the source data, the welfare survey, and applies the estimated parameters to the target census data to simulate the distribution of welfare in the full census. This requires the strong assumption that the distribution of the household characteristics used as predictors on the right-hand side of the regression, such as household size or education, are identical in the survey and census

[10] It is also the same method that the World Bank adopted for the previous poverty maps of Sri Lanka in 2002 and 2012/13.

data. This assumption will be violated if the source and survey data are collected in different time periods or if the questions are asked in significantly different ways, which could bias the poverty estimates. In addition, traditional ELL limits the set of household variables to those present in both the survey and census data. While more recent versions of ELL recommend including village level variables in addition to household level variables, to our knowledge no official poverty map has been estimated using only village level variables. Largely for this reason, poverty maps are typically produced only once every ten years, when a new census becomes available.

We use an alternative version of ELL, very similar to that described in Lange, Pape, and Wollberg (2018), that uses only community-level characteristics in the model, in this case at the GND and DSD levels.[11] This approach estimates the relationship between community level indicators in the target data and the welfare in the source data. Therefore, the exact same data is used both to predict consumption in the modeling stage and simulate it in the simulation stage, which by construction meets the strong assumption in traditional ELL that the predictors in the source and target data are drawn from the same distribution. In addition, using only village level characteristics allows the prediction model to include variables that are not present in the source data. We chose to link the survey data to the census data in this way because in this case, the source data was collected in 2016 and the target census data was collected in 2011. Not surprisingly, mean household characterstics changed significantly during those five years, making it implausible to assume that these characteristics are drawn from the same distribution (See Appendix Table 1).

The main drawback of linking the source and target data geographically is that predictors are only available at the community level are therefore unable to distinguish any variation in welfare within communities. Therefore, the resulting small area estimates will not be as precise as a traditional poverty map estimated using comparable household-level indicators present in both the source and target data. Because the linked ELL approach generates less-precise estimates than traditional ELL, applying it requires that we carefully check and consider the precision of the estimates, as measured for example by the coefficient of variation (CV).[12]

To apply this method, we start In the first stage by developing a model of household consumption per capita using data from HIES 2016. The estimation model is the following:

$$\ln(Y_i) = \beta X_c + u_{ic} \qquad (1)$$

where $ln(Y_{ic})$ is the log of per capita consumption of the i^{th} household, X_{ic} is a vector of explanatory variables at the community (GN and DS) level, and u_{ic} is the error term. A challenge in estimating equation (1) is heteroskedasticity in the error term u_{ic}, which is often prominent in household data. This is addressed by breaking the error term into two components, one at the cluster level and the other at the household level:

$$u_{ic} = \eta_d + \varepsilon_{ic} \qquad (2)$$

Where d is the DS division. Both components are assumed to be independent of the explanatory variables X_{ic} and independent of each other. However, the variance of the second component (σ_ε^2) is assumed to vary across DS Divisions. Equation (1) then becomes:

$$\ln(Y_{ic}) = \beta X_c + \eta_d + \varepsilon_{ic} \qquad (3)$$

[11] The method differs from Lange, Pape, and Wollberg (2018) by utilizing the Empirical Bayes estimation method, as described below.

[12] The CV is defined as the ratio of the standard error and the mean, often expressed as in percentage terms. It is more informative than the standard error in comparing the precision of estimates across units with different means.

Equation (3) is estimated by the Feasible Generalized Least Square method, which takes into account differences in the distribution of errors across communities. Unlike the Ordinary Least Square (OLS) method, the GLS method estimates not only the coefficients but also the variance of the error term ε_{ic}, using community characteristics. More detailed discussions on the small area estimation method can be found in Elbers, Lanjouw, and Lanjouw (2003), Van der Weide (2014), and Nguyen et al. (2018).

In the second stage, we combine the parameters estimated in the first stage and the explanatory variables in the census data to impute per capita consumption for each household in the 2012 Census. After obtaining the estimated distributions of coefficients and errors from the second step, we randomly draw coefficients and errors from these estimated distributions to simulate household expenditure for each household in the census. This simulation is repeated 100 times so that we can estimate the poverty rate for each DSD as the average poverty rate over the 100 simulations, and their standard errors as the standard deviations of the 100 simulation rounds. We use the Stata user-written command *sae* described in Nguyen et al. (2018) to carry out the modeling and simulation stages of the methodology.

An important extension to the methodology described in ELL (2003) is incorporating the Empirical Bayes (EB) prediction (van der Weide 2014, Huang and Hidiroglu 2003).[13] EB incorporates survey data to improve the precision of the location effects. This is done initially in the first modeling stage, where a modified version of GLS incorporates survey data, following Huang and Hidiroglu (2003) and Henderson (1953). These random location effects with higher precision are then incorporated into the simulation stage, so that the distribution of the location effects is obtained by utilizing observation-specific residuals of observations sampled in the location. Using the Empirical Bayes method is critical in this case because the predictor variables only vary across GNs. As a result, incorporating the survey data into the simulations through the estimated area effect makes small area estimates significantly more precise.

We conduct a cross-validation exercise to test the accuracy of our results. Cross-validation is a technique to assess how the results of an analysis will generalize to an independent dataset. We conduct cross-validation by randomly dividing the 331 DSDs from the HIES sample into ten groups. We then withhold one group from the sample, use the remaining nine groups to estimate poverty rates into the omitted part, and compare predicted poverty rate on those DSDs to the direct estimates obtained from HIES sample. This is repeated ten times, one for each DSD group. We examine the Mean Squared Error (MSE) and Mean Absolute Deviation (MAD) to evaluate our models. Lower MSE and MAD would suggest that the models perform better in out-of-sample predictions.

3b. Variable and model selection

We first identify candidate variables present in the census and remote sensing data that can be used to model household consumption. As discussed above, the model contains only community-level characteristics.[14] Based on existing literature and discussions with DCS staff, we shortlisted the following categories of explanatory variables: household size, dependency ratio, gender ratio, district dummies, housing quality (cooking fuel, lighting fuel, toilet, waste, water, roof, wall, floor), home ownership, household assets, characteristics of the household head (age, gender, marital status, employment status,

[13] See Rao and Molina (2010) on the EB methodology. Haslett (2016) discusses the strengths and weaknesses of the ELL and EB methods.
[14] The exception is district dummies, which are consistent between Census 2012 and HIES 2016.

education). We generate means at the GND and DSD levels of these variables from Census 2012 to better capture the determinants of household consumption. We also include remote sensing variables, described in Section 2c, as candidate variables for our consumption model. Since remote sensing variables are more contemporaneous with household consumption than variables from the Census, we expect them to capture shocks households may have experienced during 2012 and 2016.

We first address multicollinearity by examining the variance inflation factors (VIF). The VIF for a variable is proportional to the R^2 from the model in which all other right-hand side variables are regressed on that variable. The greater the variation in a variable that can be explained by other covariates, the higher is its VIF and the less it will add to the model's explanatory power. We drop variables whose VIF exceeds 5. We conduct this exercise sequentially for all variables until there are no variables in our model with VIF greater than 5.[15]

The specification of the consumption model affects the DSD-level poverty rates. The set of variables included in each specification was chosen from the list of candidate variables using the Least Absolute Shrinkage and Selection Operator (LASSO). The LASSO chooses a set of predictive variables while avoiding over-fitting the model to a sample (Tibshirani, 1996). This method involves regressing all candidate variables on the log of household per capita consumption. The resulting estimates include zero coefficients for some variables, which are excluded in the subsequent estimation.[16] The remaining variables with non-zero coefficients were then included in the small area estimation model, following the "post-LASSO" estimator (Belloni and Chernozhukov, 2012). Using LASSO for variable selection is inferior to alternative forward selection methods in some cases (Luo and Chen, 2014), particularly when errors are heteroscedastic (Kozbur, 2017). Nonetheless, the post-LASSO estimator is a popular, important, and easily applied generic high-dimensional estimation strategy that produces unbiased estimates of welfare. This procedure aims to reduce multicollinearity while selecting the most predictive variables.[17]

We estimate thee models for different areas of Sri Lanka to better capture geographical differences in HIES data (final specifications of the Alpha and Beta models are presented in Appendix Tables 3a-5b). This is a contrast to the 16 models that were estimated for the 2012/13 poverty map (World Bank 2015) and 26 models that were estimated for the 2002 poverty map (World Bank 2005). These areas were chosen in close consultation with DCS. The areas for which we estimate separate models of consumption include: Western Province (which includes the capital city Colombo and surrounding areas), Northern and Eastern Provinces (which were affected by conflict between the 1980s and 2009), and all other provinces. Differences in lifestyles, preferences, and consumption patterns are likely to be significant across different parts of the country. For example, ownership of vehicles may explain much variation in consumption in big cities but not so in rural areas where a small proportion of households own them. Similarly, the gender bias against female-headed households might also be stronger in rural provinces such as North Central and North Western than in Colombo. Thus, estimating separate consumption models for smaller and relatively homogenous areas is likely to produce more accurate results than estimating a single model for the entire country. However, larger regression samples also allow for more expansive model specifications, which could help improve prediction power.

---

[15] Applying this filter makes the final model more parsimonious while negligible reducing its predictive performance.
[16] We use the Stata package *lassoregress*.
[17] Stepwise selection is another popular method to select covariates. Zhao and Lanjouw (2014, p.58) suggest a rule of thumb: a small area estimation model should not include more than the square root of N variables, where N is the sample size.

4. Results and robustness checks

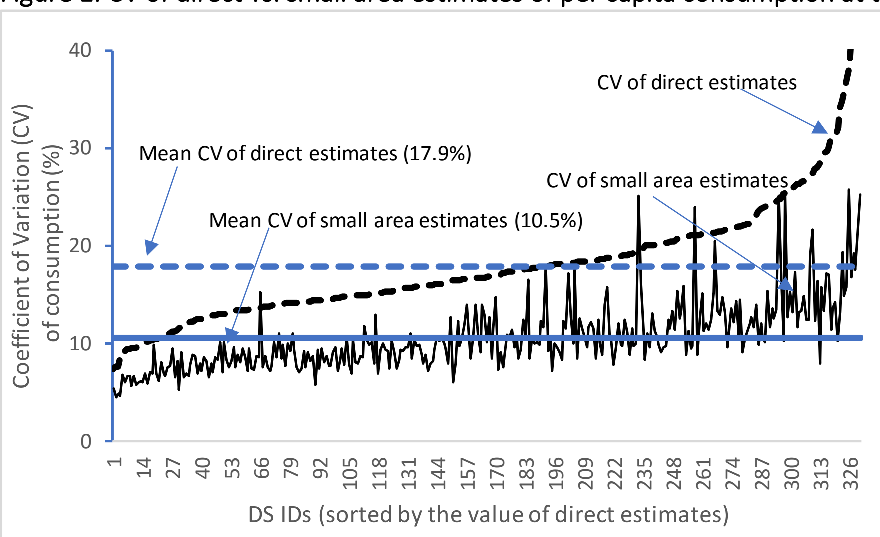
This section first discusses the reliability and accuracy of the poverty map estimates. We use the coefficient of variation and cross-validation as criteria to judge how the reliability of our estimates compare with the direct estimates obtained from HIES 2016. We also discuss any bias in our prediction of poverty rate at the district-level compared to official estimates that are representative at this level.

4a. Reliability and accuracy of estimates

Improving the precision of welfare estimates is a primary goal of any poverty mapping exercise. This is because consumption or income surveys are usually not representative at lower administrative levels because of sample size limitations. National statistical agencies often use the magnitude of CV as a criterion to decide whether to publish direct estimates at any given administrative level. The threshold level of CV often cited is 10%. Improving the precision of our estimates is especially important in our case since we estimate our consumption model with geographic aggregates of variables from the census rather than household-level variables with much larger variation.

Figure 1 presents a comparison of the CV of direct estimates of per capita consumption at the DSD-level with comparable small area estimates.[18] We see that the average CV of small area estimates (10.5%) is more than a third below the CV of direct estimates (17.9%) of per capita consumption. The improvement in CV is generally larger for DSDs with higher CV of direct estimates and is never higher than 25.7. While the average CV of per capita consumption across DSDs is not below 10%, it may be close enough to the threshold for DCS to publish them.[19]

Figure 1: CV of direct vs. small area estimates of per capita consumption at the DSD level



[18] We use the Horvitz-Thompson estimator that takes clustering into account, unlike Stata's *svy* command, which gives a mean CV of consumption of 10.8.

[19] Comparing the CV of the poverty rate, unlike the CV of consumption, is difficult since 111 out of 329 DSDs surveyed in HIES do not have any sample households that are poor at the national line, giving us a zero poverty rate, zero standard error, and an undefined CV of poverty.

A criterion for measuring the accuracy of small area estimates is how closely they line up with the survey estimates for the lowest administrative level at which the survey is representative, which in the case of HIES 2016 is districts. Appendix Table 2 presents a comparison between the poverty rate at the district level computed from HIES 2016 data, using sampling weights, and our small area estimates. We see that the average discrepancy between these two estimates is -0.1 percentage points, while the average absolute difference is 0.96 percentage points. Closer examination of the table shows that the difference in poverty rate is within one percentage point for 13 out of 25 districts. Most of the largest differences are in districts in Northern and Eastern provinces that also have the highest poverty rates.

We apply a re-scaling factor to the at the district-level poverty-estimates to ensure that our estimates match those published by DCS. Publishing small area estimates that are different from official estimates at the district level (at which HIES 2016 is representative) presents communication challenges.[20] The adjustment factor we use is simply the ratio of the direct estimates and the small area estimates at the district-level (last column of Appendix Table 2). We apply this ratio to the poverty estimates of all DSDs within a district. For example, the poverty rate of all DSDs within Colombo are scaled up by a factor of 1.13.[21] The districts with the largest and smallest rescaling factors are in the ex-conflict Northern and Eastern provinces which not only have among the highest poverty rates within the country but also saw the largest changes in poverty rates between HIES 2012/13 and HIES 2016. For example, Mannar district, whose poverty rate dropped from 20.1% in HIES 2012/13 to 1.0% in HIES 2016, has a rescaling factor of 0.35, suggesting the small area estimates are about 2.5 times higher than that of the HIES 2016 estimates. Similarly, Kilinochchi district, whose poverty rate increased from 12.7% to 18.2% has the rescaling factor of 1.51, suggesting that the small area estimates are about a third lower than the HIES 2016 estimates.

4b. Levels and changes in poverty

We next discuss the levels and changes in the poverty estimates at the DSD-level based on results from our small area estimates. Figure 2 contains two choropleth maps that allow a descriptive comparison of Sri Lanka's poverty rate at the national line estimated at the DSD-level for 2012/13 and 2016.[22] This is the first set of poverty maps that allow comparison of poverty in all DSDs of Sri Lanka at two points in time.[23] The DSDs in both maps are grouped into the following bins of poverty rate: 0-10%, 10-20%, 20-30%, and 30-50%. The estimated poverty rate at the DSD level ranges from 0.6% to 45.1% in 2012/13 and 0.1% to 40.4% in 2016. Darker areas on the maps denote DSDs with higher poverty rates whereas the lighter areas denote DSDs with lower poverty rates.

[20] We recognize that HIES itself is based on a single sample from the population and therefore yields estimates with some margin of error. However, the previous poverty maps for Sri Lanka did not include this adjustment and resulted in some confusion in the media. DCS also feels strongly about including this adjustment.

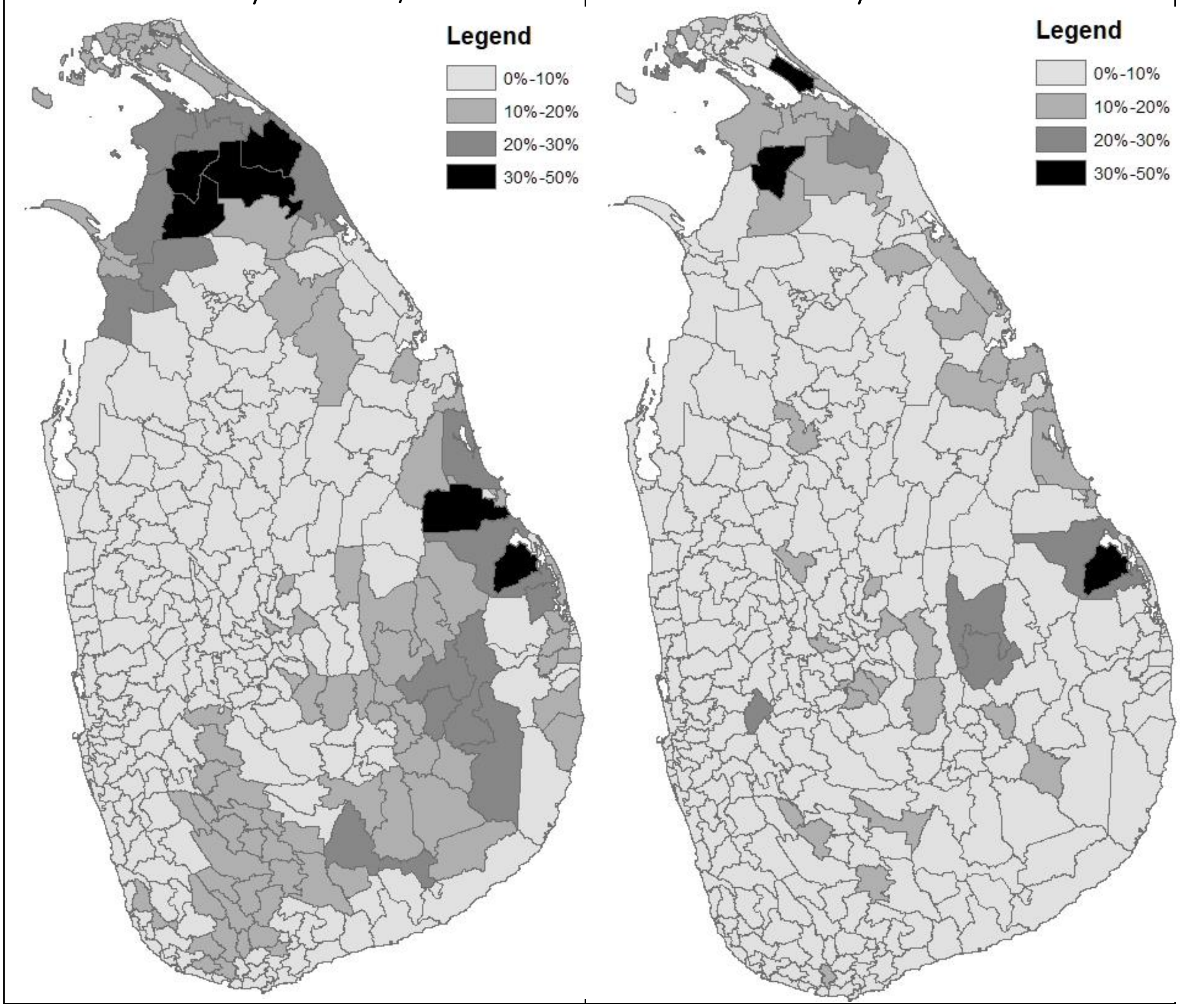
[21] We assume that the CV of per capita consumption remains the same even after the poverty rate is rescaled, which may not necessarily be a reasonable assumption.

[22] A strict comparison may not be valid due to the minor differences in methodology used to generate the small area estimates behind these two maps.

[23] The 2002 poverty map did not cover the Northern and Eastern parts of the country.

Figure 2: DSD-level poverty rate in 2012/13 vs. 2016 from Small Area Estimation

a. Poverty rate in 2012/13	b. Poverty rate in 2016



We can see that the poverty rate is below 10% for a large part of the country. We also see that the 2016 map has more DSDs with poverty rates below 10%. This is consistent with the progress that Sri Lanka has achieved in alleviating poverty and that the poverty rate at the national line decreased from 6.7% in 2012/13 to 4.1% in 2016. A majority of DSDs in Colombo and Gampaha districts, as well as sizable parts of Hambantota and Mannar districts are particularly well-off, with estimated poverty rates below 2%. High poverty incidence is concentrated in Kilinochchi, Mullaitivu, Batticaloa, Jaffna, and Trincomalee districts in the Northern and Eastern provinces. The map also reveals significant geographical disparity among DSDs in some districts. Poverty rates in DSDs in Batticaloa, for example, vary widely from 0.9% to 30.7%. In Mullaitivu, DSD poverty rates range from 4.3% to 40.4%.

Consistent with the widespread drop in poverty rates, we see that the pockets of poverty largely shrank between 2012/13 and 2016. Although the Northern and Eastern provinces experienced large reductions in poverty during 2012/13 and 2016, these provinces still have the DSDs with the highest poverty rates within the country. The high rates of poverty in these areas were to be expected, given that they lay at the center of the civil conflict for more than 30 years. In 2012/13, the two provinces were at the beginning of a rapid process of resettlement and economic rehabilitation, which have been funded by many local and outside sources. Since then, development programs as well as continued economic growth may have considerably improved conditions in these areas. Another remaining pocket of poverty is in the south-central provinces of Uva and Sabaragamuwa, although this region has also seen widespread poverty reduction since 2012/13.[24]

Appendix Table 3 lists the DSDs with the highest poverty rates within the country, their poverty rates at the national line, the CV of poverty, and the CV of per capita consumption. Consistent with the description earlier, we see that the many of poorest DSDs are in the Northern and Eastern provinces, including in Mullaitivu, Kilinochchi, and Batticaloa districts. We, however, see little overlap between the poorest DSDs in 2012/13 and 2016; only 4 of the poorest 10 DSDs in 2012/13 are also among the poorest 10 DSDs in 2016. Only 5 of the poorest 20 DSDs in 2012/13 are also among the poorest 20 DSDs in 2016. Many of the poorest DSDs in 2012/13 saw large reductions in poverty between 2012/13 and 2016; the average poverty reduction was 15 percentage points for this group during this period. This may be a testament to the fact that government policies informed by the previous poverty map may have been successful in alleviating poverty.

Figure 3: Scatter-plot of poverty in 2012/13 vs. change in poverty between 2012/13 and 2016

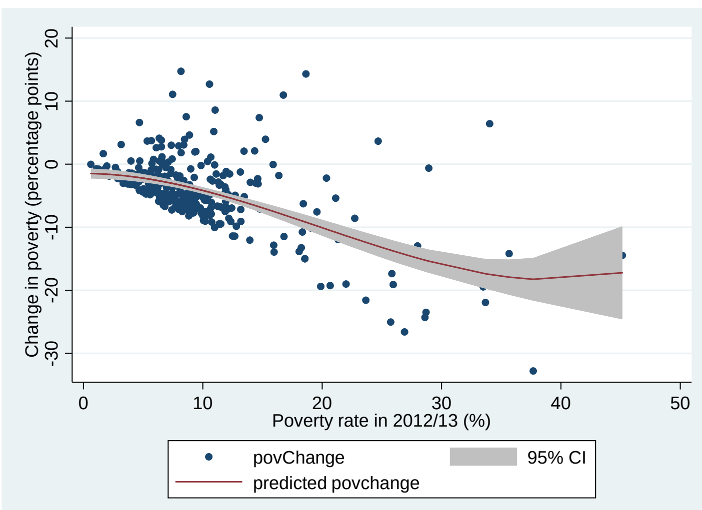


Figure 3 presents a scatter-plot of the change in poverty rate, in percentage points, between 2012/13 and 2016 on the y-axis against the poverty rate in 2012/13 at the DSD-level on the x-axis. We see a negative relationship between these two variables, although there is significant heterogeneity for DSDs with high poverty rate in 2012/13. A likely explanation in this heteroscedasticity is measurement error in either or both poverty maps. Despite this heterogeneity, we see that a majority of poorest DSDs saw reduction in poverty between 2012/13 and 2016.[25] In about 12% of the DSDs, the poverty rate in 2016 increased by

[24] Future analysis could examine correlates of changes in poverty rate between 2012/13 and 2016 at the DSD level. Such analysis could include census variables such as access to infrastructure or population density (urbanization).

[25] We could also examine the Growth Incidence Curve of per capita consumption in follow-up work.

more. This is not completely inconsistent with the fact that three out of 25 districts had a higher poverty rate in 2016 than in 2012/13.[26]

4c. Robustness of results

The results described in sections 4a and 4b are a result of a variety of assumptions and modeling decisions. This section describes the results of the following checks we conducted to ensure that the results described earlier are robust: cross-validation, alternative covariate sets in the consumption model, alternative poverty line, and national vs. regional consumption models.[27]

Tables 1 and 2 present the results of the robustness checks. The columns of these tables represent different sets of covariates used as candidate variables in the consumption model. In addition to variables derived from the census, we ran estimation with only remote sensing variables as well as both variables from census and remote sensing. As discussed in Section 3, variables derived from the census and remote sensing both have their own strengths and weaknesses. Although much of our analysis is conducted at the national poverty line (which corresponds to approximately $2.48 PPP/day), we also examine results at the World Bank's Lower Middle-Income Country Line of $3.20 PPP/day. We finally compare results from a single national model of consumption to three separate models for three areas of the country: Western province; Northern and Eastern province; and all other provinces.[28]

Table 1 presents the results from a 10-fold cross-validation exercise to assess how the results of our analysis will generalize to an independent dataset. The statistics we examine are the Mean Squared Error (MSE) and Mean Absolute Deviation (MAD). We surprisingly do not find large differences in MSE and MAD across the three covariate sets in our analysis, either at the national poverty line or the LMIC poverty line. We also do not find large differences in MSE and MAD between simulations done with a national model or three regional models of consumption. We do however, find that MSE and MAD are larger for simulations using the LMIC poverty line than the national line. This makes sense since both MSE and MAD are measures of absolute error, which we would also expect to be higher for higher poverty rates.

[26] In fact, the DSD with the second largest increase in poverty between 2012/13 and 2016 is Pachchilaipalli DSD in Kilinochchi district. The poverty rate of Kilinochchi district increased from 12.7% in 2012/13 to 18.2% according to HIES data.

[27] We also conducted estimation using the Fay-Herriot, BHF, and HBSAE methods. We will discuss these results in a future version of this report.

[28] We also attempted separate models for the rural, urban, and estate sectors. Results were qualitatively similar to that those from three regional models. However, we believe that the regional models better capture economic heterogeneity within the country than sectoral models. We did not attempt to estimate 16 models as was done for the 2012/13 poverty map or 26 models as was done for the 2002 poverty map since the large number of candidate variables in our estimation models meant that the degree of freedom in our estimation models would shrink significantly.

Table 1: Cross-validation results

District	National poverty line (Rs. 4,166/month)			
	Source of covariates:			
	Census (HH & area-level vars)	Census (area-level vars)	Remote Sensing	Census + Remote Sensing
National Model				
Mean Squared Error	0.246	0.176	0.175	0.178
Mean Absolute Deviation	3.231	2.485	2.583	2.533
Regional Models				
Mean Squared Error	0.261	0.182	0.180	0.179
Mean Absolute Deviation	3.373	2.525	2.664	2.544
	Lower Middle-Income Country line ($3.20 PPP/month)			
National Model				
Mean Squared Error	0.952	0.484	0.495	0.486
Mean Absolute Deviation	7.405	5.260	5.328	5.204
Regional Models				
Mean Squared Error	1.030	0.518	0.517	0.507
Mean Absolute Deviation	7.735	5.364	5.500	5.270

Table 2 presents the national poverty rate and the coefficient of variation of poverty as well as per capita consumption for estimations conducted with three different covariate sets. This table also presents results for the national poverty line and the LMIC poverty line. We see that the estimated poverty rate at the national line (4.2%) almost exactly matches the official estimate (4.1%) for the specification with census and remote sensing variables. It is reassuring that the poverty rates estimated by the specifications with census or remote sensing variables are not too far off from official estimate either. The estimated poverty rate at the LMIC line for all the specifications are very similar to each-other, but below the official estimate of 9.7%. Although it is reassuring to see that the bias in prediction is not large irrespective of the specification we choose, the fact that we rescale poverty estimates to match official estimates at the district level means that this is even less of a concern.

We see larger differences in the mean CV of poverty and consumption across different specifications and poverty lines (Table 2). We see that the specifications with only census variables generates results that have the lowest CV. The specifications with only remote sensing variables generate results with the highest CV, while the specifications with both the census and remote sensing covariates generate results with CV somewhere in the middle. The CV of poverty is much lower when we examine results at the LMIC poverty line which yields a poverty rate that is more than twice that at the national poverty line.

Table 2: Mean and CV of simulated consumption and poverty rates

	National poverty line (Rs. 4,166/month)				
		Source of covariates:			
	HIES 2016	Census (HH & area-level vars)	Census (area-level vars)	Remote Sensing	Census + Remote Sensing
National model					
Poverty rate at the national line (%)	4.1	5.5	4.2	4.2	4.1
Mean CV of consumption across DSDs (%)	17.9	8.2	9.7	10.9	10.5
Mean CV of poverty across DSDs (%)	64.6	35.5	42.8	50.1	46.5
Regional models+					
National poverty rate (%)	4.3	5.7	4.2	4.4	4.2
Mean CV of consumption across DSDs (%)	17.9	8.2	9.7	10.9	10.4
Mean CV of poverty across DSDs (%)	64.6	35.1	43.7	50.0	45.5
	LMIC line ($3.20 PPP/day)*				
National model					
Mean CV of consumption across DSDs (%)	50.1	28.2	34.7	40.0	37.5
Mean CV of poverty across DSDs (%)	9.7	10.8	8.7	8.6	8.4
Regional models+					
National poverty rate (%)					
Mean CV of consumption across DSDs (%)	50.1	28.0	35.1	39.8	36.7
Mean CV of poverty across DSDs (%)	9.7	11.2	8.7	8.8	8.6

* LMIC stands for Lower Middle-Income Country (World Bank definition).

+ Regional models estimated for Western; Northern & Eastern; Other provinces.

5. Conclusions

Small area estimation techniques borrow strength from auxiliary data with comprehensive coverage to generate more precise estimates of poverty for small geographic areas or demographic groups. Traditionally, this exercise relies on estimating a model using household characteristics in a welfare survey and applying the estimated coefficients and parameters to census data to simulate poverty rates. A key drawback of this method is that the household characteristics used to link the census and survey data must be drawn from the same underlying probability distribution. This implies that the census and the survey need to be conducted around the same point in time. The longer the gap between the census and the survey, the more difficult it is to justify this assumption and the greater the likelihood the estimates will be biased relative to true poverty at the time of the survey. Therefore, until recently it has not been possible to monitor changes in poverty in small areas except when a new census is available. We present a poverty map for Sri Lanka in 2016 using data from a household survey in 2016, data from satellite imagery taken in 2017 and 2018, and the 2012 Census. The main challenge was the temporal gap between the HIES 2016 and Census 2012, which we addressed by excluding household characteristics variables from the consumption model.

Satellite imagery is an exciting new source of data for the estimation of geographically-disaggregated poverty statistics on a frequent basis. The availability of satellite imagery and the ability to process it has improved significantly in recent years. Satellite imagery has extensive spatial and temporal coverage, especially for recent years, and captures a significant portion of spatial differences in living standards across communities. We apply a variety of features derived from both satellite imagery and the 2012 census, linked with the 2016 household survey at the community level, to generate subdistrict poverty estimates for Sri Lanka in 2016. Our empirical model links household per capita consumption to community characteristics using HIES 2016. Community characteristics are selected from a pool of census and satellite indicators using the LASSO to avoid overfitting. The selected variables are then employed in a simulation framework that incorporates random location effects, a GLS heteroscedasticity correction, and Empirical Bayes estimation.

Incorporating geographically linked satellite indicators reduces the average coefficient of variation of subdistrict poverty estimates by approximately one third compared with direct survey-based estimates. This is roughly equivalent to doubling the effective size of the sample. The coefficient of variation of augmented subdistrict estimates of mean per capita consumption is 10.3%. This is close to the survey's 10% percent target for district level estimates, which are routinely published. The analysis compares the performance of the estimates derived from satellite indicators with the traditional method of applying a model estimated on 2016 survey data to 2012 census data. Not surprisingly, predictions derived from satellite and census data generate much more plausible estimates of changes in subdistrict poverty rates. Models estimated using either census data alone, or satellite data alone, offer similar predictive performance to models that draw on both types of variables. Overall, the results strongly suggest that indicators derived from satellite imagery or other geographically linked data should routinely be combined with survey data to better track changes in poverty at local areas, even in the absence of a new census. Future work could examine characteristics that help explain the substantial heterogeneity in poverty reduction across DSDs, particularly among the DSDs that were the poorest in the country in 2012/13.

References

Belloni, Alexandre, and Victor Chernozhukov. "Least squares after model selection in high-dimensional sparse models." Bernoulli 19.2 (2013): 521-547.

Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." international Conference on computer vision & Pattern Recognition (CVPR'05). Vol. 1. IEEE Computer Society, 2005.

Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw. "Micro–level estimation of poverty and inequality." Econometrica 71.1 (2003): 355-364.

Engstrom, Ryan, Vidhya Soundararajan, and David Newhouse. 2019, "Estimating Small Area Population Density Using Survey Data and Satellite Imagery: An Application to Sri Lanka", Policy Research Working Paper series.

Engstrom, Ryan, Jonathan Hersh, and David Newhouse. "Poverty from space: using high-resolution satellite imagery for estimating economic well-being." World Bank Policy Research Paper no. 8284 (2017)

Funk, Chris, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell & Joel Michaelsen.
"The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes". Scientific Data 2, 150066. doi:10.1038/sdata.2015.66 2015.

Hansen, Matthew C., Peter V. Potapov, Rebecca Moore, Matt Hancher, S. A. A. Turubanova, Alexandra Tyukavina, David Thau et al. "High-resolution global maps of 21st-century forest cover change." science 342, no. 6160 (2013): 850-853.

Henderson, C. R. (1953). Estimation of variance and covariance components. Biometrics 9 (2), 226–252.

Huang, R. and Hidiroglou, M. (2003). Design consistent estimators for a mixed linear model on survey data. Joint Statistical Meetings, Section on Survey Research Methods 1897–1904.

Huang, Xin, Zhang, Liangpei, and Li, Pingxiang (2007) Classification and Extraction of Spatial Features in Urban Areas Using High-Resolution Multispectral Imagery. IEEE Geoscience and Remote Sensing Letters, 4(2).

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794.

Kozbur, Damian. "Testing-based forward model selection." American Economic Review 107.5 (2017): 266-69.

Lange, Simon, Utz Johann Pape, and Peter Pütz. Small area estimation of poverty under structural change. The World Bank, 2018.

Luo, Shan, and Zehua Chen. "Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space." Journal of the American Statistical Association 109.507 (2014): 1229-1240.

Mehrotra, Rajiv, Kameswara Rao Namuduri, and Nagarajan Ranganathan. "Gabor filter-based edge detection." Pattern recognition 25.12 (1992): 1479-1494.

Myint, Soe W, Mesev, Victor, and Lam, Nina (2006) Urban Textural Analysis from Remote Sensor Data: Lacunarity Measurements Based on the Differential Box Counting Method. Geographical Analysis, ISSN 0016-7363

Nguyen, M. C., Corral, P., Azevedo, J. P., & Zhao, Q. (2018). Sae: A Stata Package For Unit Level Small Area Estimation. The World Bank.

Pesaresi, Martino, Gerhardinger, Andrea, and Kayitakire, François (2008) A Robust Built-Up Area Presence Index by Anisotropic Rotation-Invariant Textural Measure. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1(3).

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-288.

Ünsalan, Cem and Boyer, KL (2004) Classifying Land Development in High-Resolution Satellite Imagery Using Hybrid Structural–Multispectral Features. IEEE Transactions on Geoscience and Remote Sensing, 42(12)

Van der Weide, Roy. GLS estimation and empirical bayes prediction for linear mixed models with Heteroskedasticity and sampling weights: a background study for the POVMAP project. The World Bank, 2014.

World Bank. 2005. A poverty map for Sri Lanka -- lessons and findings : policy note (English). Washington, DC: World Bank. http://documents.worldbank.org/curated/en/129331468113649429/A-poverty-map-for-Sri-Lanka-lessons-and-findings-policy-note

Zhao, Qinghua, and Peter Lanjouw. "Using PovMap2-A User’s Guide." The World Bank (2009).

Appendix

Appendix Table 1: Summary statistics of variables from Census 2012 and HIES 2016

	Census 2012				HIES 2016			
Variables	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
Household size	3.77	1.64	1	30	4.46	1.63	1	13
Age of household head in years	49.87	14.51	12	99	52.30	13.73	14	99
Head of household is a male	0.76	0.43	0	1	0.78	0.42	0	1
Household head is married	0.85	0.36	0	1	0.82	0.38	0	1
Household head is unmarried	0.04	0.19	0	1	0.01	0.11	0	1
Household head is married	0.85	0.36	0	1	0.82	0.38	0	1
Household head is widowed	0.10	0.30	0	1	0.14	0.35	0	1
Household head is divorced/separated	0.02	0.13	0	1	0.02	0.16	0	1
Household head's schooling: None	0.04	0.20	0	1	0.03	0.18	0	1
Household head's schooling: Less than Grade 5	0.22	0.42	0	1	0.21	0.41	0	1
Household head's schooling: Grades 6-10	0.41	0.49	0	1	0.48	0.50	0	1
Household head's schooling: GCE O/A levels	0.29	0.45	0	1	0.24	0.43	0	1
Household head's schooling: Degree and higher	0.03	0.18	0	1	0.03	0.17	0	1
Household head's status: Unemployed	0.01	0.07	0	1	0.01	0.08	0	1
Household head's status: Employed in public sector	0.10	0.30	0	1	0.10	0.29	0	1
Household head's status: Employed in private sector	0.57	0.50	0	1	0.62	0.48	0	1
Household head's status: Not in labor force	0.33	0.47	0	1	0.28	0.45	0	1
Highest education in household: None	0.01	0.10	0	1	0.00	0.06	0	1
Highest education in household: Less than Grade 5	0.06	0.23	0	1	0.03	0.16	0	1
Highest education in household: Grades 6-10	0.37	0.48	0	1	0.38	0.49	0	1
Highest education in household: Less than GCE O/A lev	0.49	0.50	0	1	0.50	0.50	0	1
Highest education in household: Degree and higher	0.07	0.26	0	1	0.08	0.28	0	1
HH's main cooking fuel is fire wood	0.78	0.41	0	1	0.69	0.46	0	1
HH's main type of lighting fuel is electricity	0.87	0.34	0	1	0.98	0.15	0	1
HH has access to a private toilet	0.87	0.34	0	1	0.93	0.26	0	1
HH has access to a water-sealed toilet	0.95	0.22	0	1	0.98	0.15	0	1
HH has access to waste disposal service by local author	0.20	0.40	0	1	0.22	0.42	0	1
HH's main source of drinking water is safe	0.87	0.34	0	1	0.85	0.36	0	1
HH owns the house it lives in	0.83	0.38	0	1	0.89	0.31	0	1
Wall made from permanent/semi-permanent materials	0.89	0.31	0	1	0.93	0.25	0	1
Floor made from permanent/semi-permanent material	0.91	0.28	0	1	0.96	0.19	0	1
Roof made from permanent/semi-permanent materials	0.98	0.15	0	1	0.92	0.27	0	1
HH owns a radio	0.70	0.46	0	1	0.64	0.48	0	1
HH owns a TV	0.79	0.40	0	1	0.89	0.31	0	1
HH owns a land phone	0.44	0.50	0	1	0.30	0.46	0	1
HH owns a mobile phone	0.80	0.40	0	1	0.92	0.28	0	1
HH owns a desktop or laptop computer	0.19	0.40	0	1	0.22	0.41	0	1

Note: HIES 2016 estimates were computed using sampling weights.

Appendix Table 2: Discrepancy between direct estimates and small area estimates of poverty rate (%)

District	Direct estimates (%)	Small area estimates (%)	Difference (% points)	Rescaling factor
Colombo	0.9	0.8	-0.1	1.13
Gampaha	2.0	1.9	-0.1	1.05
Kalutara	2.9	4.3	1.4	0.67
Kandy	5.5	5.1	-0.4	1.08
Matale	3.9	4.0	0.1	0.98
Nuwara Eliya	6.3	6.3	0.0	1.00
Galle	2.9	3.9	1.0	0.74
Matara	4.4	5.0	0.6	0.88
Hambantota	1.2	2.7	1.5	0.44
Jaffna	7.7	5.9	-1.8	1.31
Mannar	1.0	2.8	1.8	0.36
Vavuniya	2.0	1.9	-0.1	1.05
Mullaitivu	12.7	11.1	-1.6	1.14
Kilinochchi	18.2	12.1	-6.1	1.50
Batticaloa	11.3	9.3	-2.0	1.22
Ampara	2.6	3.8	1.2	0.68
Tricomalee	10.0	6.9	-3.1	1.45
Kurunegala	2.9	2.8	-0.1	1.04
Puttlam	2.1	3.1	1.0	0.68
Anuradhapura	3.8	3.1	-0.7	1.23
Polonnaruwa	2.2	4.1	1.9	0.54
Badulla	6.8	9.3	2.5	0.73
Moneragala	5.8	6.4	0.6	0.91
Ratnapura	6.5	7.4	0.9	0.88
Kegalle	7.1	5.5	-1.6	1.29
Average			**-0.1**	**0.96**

Appendix Table 4a: Beta Model for Western Province

	Coeff.	T-stat
Mean household size in GN	-0.299***	(-7.85)
Mean age of household heads in GN	-0.00566	(-0.87)
Share of male household heads in GN	0.606*	(1.95)
Share of household heads that are not in the labor force in GN	0.569**	(2.22)
Share of households with highest education = Grades 6-9 in GN	-0.190	(-1.45)
Share of households with highest education of Degree or higher in GN	1.770***	(6.79)
Percent of HHs in GN with water-sealed toilet	0.00123	(0.76)
Share of households whose walls are made from perm/semi-perm materials in GN	0.260	(1.44)
Share of households that own a radio in GN	0.323**	(2.05)
Share of household heads that are employed in public sector in DS	-2.703**	(-2.26)
Mean household dependency ratio (children<15) in DS	-1.890**	(-2.35)
Mean gender ratio in DS	2.215**	(2.25)
Share of households whose main lighting fuel is electricity in DS	0.590	(0.68)
GN population, millions	8.229*	(1.89)
Tree Cover - Loss, GN	2.123**	(2.07)
Z-score of 2016-Q1 rainfall in GN	0.299	(1.23)
Z-score of 2016-Q2 rainfall in GN	-0.0375	(-0.87)
Z-score of 2016-Q3 rainfall in GN	-0.255***	(-3.03)
hog_sc5_mean_gn	23.78**	(2.05)
sfs_sc3_w_mean_gn	0.0694	(1.11)
sfs_sc5_std_gn	-0.0216	(-0.99)
sfs_sc7_mean_gn	-0.00322	(-0.40)
sfs_sc7_std_gn	-0.0336*	(-1.79)
Built up area in 1990, DS	-0.00271*	(-1.89)
Sector = Rural	0.0339	(0.95)
Sector = Estate	-0.266***	(-3.21)
Constant	5.157***	(3.08)
R-squared	0.18	
N	4874	

t statistics in parentheses. * p<.10, ** p<.05, *** p<.01

Appendix Table 4b: Beta Model for Northern and Eastern Provinces

	Coeff.	T-stat
Mean household size in GN	0.0115	(0.26)
Mean age of household heads in GN	-0.00175	(-0.28)
Share of household heads that are employed in public sector in GN	0.475**	(2.27)
Share of household heads that are not in the labor force in GN	0.183	(1.17)
Share of households with highest education = Grades 6-9 in GN	-0.515***	(-3.98)
Percent of HHs in GN with water-sealed toilet	0.000934	(1.33)
Percent of HHs in GN with permanent roof	-0.00961***	(-2.78)
Share of households that own the house they live in in GN	-0.145	(-1.18)
Share of households that own a radio in GN	0.163	(1.57)
Share of households that own a mobile phone in GN	0.262*	(1.93)
Percent of HHs in GN with access to internet	0.00620***	(3.64)
Share of household heads that are employed in public sector in DS	0.0829	(0.14)
Share of household heads that are not in the labor force in DS	0.778	(1.34)
Mean household dependency ratio in DS	-0.665	(-1.28)
Percent of HHs in DS with permanent roof	0.00185	(0.19)
Share of households whose walls are made from perm/semi-perm materials in DS	0.0650	(0.28)
GN area, sq km X 100000)	1.52e-09	(0.56)
Mean night light intensity in 2016, GN	-0.0000309	(-0.03)
GN population, millions	13.00	(1.30)
Mean elevation of GN	-0.000426	(-0.47)
Tree Cover - Gain, GN	0.437	(0.66)
Tree Cover - Loss, GN	-13.60**	(-2.49)
Built up area in 2014, GN	0.102*	(1.75)
Built up area in 2000, GN	0.0290	(0.33)
Built up area in 1990, GN	0.103	(1.63)
Squared Z-score of 2016-Q2 rainfall in GN	0.000378	(0.65)
Squared Z-score of 2016-Q3 rainfall in GN	0.0701	(1.54)
fourier_sc7_std_gn	-0.00116	(-0.13)
gabor_sc7_std_gn	0.0425	(1.34)
lsr_sc3_line_std_gn	0.0633	(0.90)
ndvi_sc3_mean_gn	-0.252	(-1.31)
pantex_sc7_min_std_gn	-0.762	(-1.12)
sfs_sc3_std_gn	-0.0849**	(-2.24)
sfs_sc3_w_std_gn	0.0378	(0.97)
sfs_sc7_std_gn	0.000220	(0.01)
Tree Cover - Loss, DS	1.219	(1.30)
Built up area in 2000, DS	0.0105	(0.77)
Built up area in 1975, DS	-0.00441	(-0.78)
Z-score of 2016-Q4 rainfall in DS	0.187	(0.99)
hog_sc7_std_ds	-3.026	(-0.74)
lac_sc7_lac_mean_ds	-0.0000110	(-0.26)
lsr_sc3_line_mean_ds	0.00817	(0.08)
pantex_sc3_min_std_ds	0.874	(0.60)
Sector = Rural	0.0715	(1.57)
Constant	9.460***	(14.25)
R-squared	0.15	
N	3201	

t statistics in parentheses. * p<.10, ** p<.05, *** p<.01

Appendix Table 4c: Beta Model for Provinces other than Western, Norther, and Eastern

	Coeff.	T-stat
Mean household size in GN	-0.144***	(-4.08)
Mean age of household heads in GN	-0.0141***	(-3.76)
Share of male household heads in GN	-0.459***	(-2.83)
Share of households with highest education = Grades 6-9 in GN	-0.269***	(-2.90)
Share of households with highest education of Degree or higher in GN	1.567***	(5.76)
Mean household dependency ratio in GN	-0.413***	(-3.60)
Percent of HHs in GN with water-sealed toilet	0.000288	(0.42)
Percent of HHs in GN with access to pipe water within premises	0.000481	(1.57)
Percent of HHs in GN with permanent roof	0.00378	(1.59)
Share of households whose floors are made from perm/semi-perm materials in GN	0.171	(1.55)
Share of households that own the house they live in in GN	-0.185***	(-3.63)
Share of households that own a land phone in GN	0.146**	(2.25)
Share of households that own a mobile phone in GN	-0.00902	(-0.09)
Percent of HHs in GN with access to internet	0.00357***	(2.58)
Share of household heads that are employed in public sector in DS	0.306	(0.64)
Share of households with highest education of less than Grade 5 in DS	-0.591	(-0.64)
Mean household dependency ratio in DS	0.400	(0.91)
Percent of HHs in DS with water-sealed toilet	0.00930*	(1.70)
Percent of HHs in DS with access to waste disposal by local authorities	0.00154	(0.71)
Share of households whose walls are made from perm/semi-perm materials in DS	-0.299	(-0.94)
GN area, sq km X 100000)	1.73e-09	(0.57)
Mean night light intensity in 2016, GN	0.00146	(1.49)
Mean slope of GN	-0.00461*	(-1.81)
Tree Cover - Gain, GN	0.661*	(1.92)
Tree Cover - Loss, GN	-0.360	(-0.70)
Built up area in 2014, GN	0.00434	(0.16)
Built up area in 2000, GN	0.000637	(0.02)
Z-score of 2016-Q1 rainfall in GN	0.286**	(2.03)
Squared Z-score of 2016-Q2 rainfall in GN	0.0000186	(0.03)
fourier_sc5_std_gn	0.00319	(0.52)
hog_sc7_mean_gn	13.68	(1.28)
lsr_sc3_line_std_gn	-0.00431	(-0.08)
ndvi_sc3_mean_gn	0.123	(0.75)
pantex_sc7_min_mean_gn	0.293	(1.25)
sfs_sc3_w_mean_gn	0.0161	(0.37)
sfs_sc5_std_gn	0.0271	(1.64)
sfs_sc7_mean_gn	0.00855**	(2.02)
Built up area in 2014, DS	-0.00245	(-1.02)
Built up area in 2000, DS	0.0101**	(2.51)
Built up area in 1990, DS	-0.000372	(-0.24)
Mean rainfall in DS (1981-2015)	-0.0000121	(-0.86)
gabor_sc3_std_ds	0.141**	(2.33)
hog_sc7_mean_ds	-22.25	(-0.64)
lsr_sc5_line_mean_ds	-0.325***	(-2.78)
lsr_sc7_line_std_ds	0.492**	(2.03)
ndvi_sc3_std_ds	-1.357**	(-2.18)
pantex_sc3_min_mean_ds	-1.367	(-0.52)
pantex_sc5_min_std_ds	-0.194	(-0.20)
sfs_sc5_std_ds	0.151	(1.37)
sfs_sc5_w_mean_ds	-0.0741	(-0.75)
Sector = Rural	0.0220	(0.62)

Sector = Estate	-0.198***	(-4.36)
District = Matara	0.00228	(0.04)
District = Hambantota	0.240***	(3.07)
District = Kurunegala	0.140**	(2.06)
District = Puttalam	0.138	(1.55)
District = Anuradhapura	0.237***	(2.89)
District = Polonnaruwa	0.162	(1.41)
District = Badulla	-0.0527	(-0.70)
District = Monaragala	0.115	(1.35)
Constant	10.44***	(2.65)
R-squared	0.15	
N	11583	

t statistics in parentheses. * p<.10, ** p<.05, *** p<.01