Policy Research Working Paper 9383

World Development Report 2021

Background Paper

Small Area Estimation of Non-Monetary Poverty with Geospatial Data

Takaaki Masaki
David Newhouse
Ani Rudra Silwal
Adane Bedada
Ryan Engstrom

WORLD BANK GROUP
Poverty and Equity Global Practice
September 2020

Policy Research Working Paper 9383

Abstract

This paper uses data from Sri Lanka and Tanzania to evaluate the benefits of combining household surveys with geographically comprehensive geospatial indicators to generate small area estimates of non-monetary poverty. The preferred estimates are generated by utilizing subarea-level geospatial indicators in a household-level empirical best predictor mixed model with a normalized welfare measure. Mean squared errors are estimated using a parametric bootstrap procedure. The resulting estimates are highly correlated with non-monetary poverty calculated from the full census in both countries, and the gain in precision is comparable to increasing the size of the sample by a factor of three in Sri Lanka and five in Tanzania. The empirical best predictor model moderately underestimates uncertainty, but coverage rates are similar to standard survey-based estimates that assume independent outcomes across clusters. A variety of checks, including adding noise to the welfare measure and model-based and design-based simulations, confirm that the main results are robust. The results demonstrate that combining household survey data with subarea-level geospatial indicators can greatly increase the precision of survey estimates of non-monetary poverty at comparatively low cost.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dnewhouse@worldbank.org.

Small Area Estimation of Non-Monetary Poverty with Geospatial Data

Takaaki Masaki

David Newhouse

Ani Rudra Silwal

Adane Bedada

Ryan Engstrom

This paper is a product of the Poverty and Equity global practice and a background report for the 2021 World Development Report. We are grateful to Nadia Belghith, Carlos Castelan, Robert Cull, Andrew Dabalen, Kristen Himelein, Dean Joliffe, Pierella Paci, Carolina Sanchez-Paramo, and Thomas Walker for their support. We thank Paul Corral, Kristen Himelein, Peter Lanjouw, and William Seitz for insightful comments and suggestions, and Partha Lahiri, Carl Morris and Roy van der Weide for helpful discussions. Finally, we thank Albina Chuwa, Dilhanie Deepawansa, Keith Garrett, Amara Satharasinghe, and Elizabeth Talbert for their help in obtaining data. All remaining errors are ours.

Introduction

The proliferation of big data obtained from satellites and mobile devices, as well as research demonstrating that satellite and mobile phone records are strongly correlated with household welfare, has sparked great interest in statistical methods that combine big data with survey data.[1] A key motivation for complementing survey data with comprehensive big data is the potential for small area estimation, which can produce more precise and granular estimates of socioeconomic indicators by combining household-level survey data in surveyed enumeration areas with predicted values from all enumeration areas, whether sampled or not. Established methods for small area estimation typically utilize household survey data and contemporaneous household census data, by first estimating a prediction model of welfare in the survey data and then using the estimated model parameters to simulate welfare in the census.[2] Because traditional small area estimation links household survey with census auxiliary data through the estimated model parameters, the set of potential predictor variables is limited to those present in both the survey and the census. Furthermore, when the wording of questions or timing of the data collection differs significantly between the census and the survey, small area estimates using the standard method can suffer from bias. Finally, because this method relies on census data, a new census is required to credibly estimate welfare changes for small areas. Census data are typically dated and collected once per decade at best, which presents a significant hurdle to deriving timely small area poverty estimates.

To overcome these challenges, we link survey data with geospatial and remote-sensing auxiliary data at the *subarea* level.[3] S*ubarea refers to* the lowest geographic level at which spatial auxiliary data can be merged with household surveys containing data on welfare and poverty, while *area* refers to the target level for the small area poverty estimates. In recent household surveys, GPS coordinates of households, enumeration areas, or villages are often available, which allows analysts to link survey data with other auxiliary sources of remote-sensing or geospatial data and improve the precision of poverty estimates.

Although linking surveys with geospatial auxiliary data generates less precise small area estimates than using a census, it offers several important benefits. Geographic linking allows the relationship between the spatial auxiliary indicators and economic welfare to be estimated directly in the prediction model, avoiding the restrictive assumption that predictors must be identically distributed in the survey and auxiliary data. In addition, some spatial auxiliary data are collected continuously and are highly predictive of subarea poverty, and both the quality and availability of geospatial data are improving rapidly. Compared to other forms of "big data" such as mobile phone data, indicators derived from satellite imagery are attractive because of the comparative ease of acquiring data and the absence of selection bias.

Despite the potential value of this approach, only a couple of existing studies have quantified the gain in the precision of small area estimates achieved by supplementing survey data with geospatial data, in the

[1] Examples of papers that find strong correlations between big data and welfare are Yeh et al (2020), Jean et al, 2016, Steele et al, 2018, Engstrom et al (2017) and Pokhriyal and Jacques (2017), among others.
[2] Elbers, Lanjouw, and Lanjouw (2003), Tarozzi and Deaton (2009), and Molina and Rao (2010), among many others.
[3] Lange, Pape, and Putz (2018).

context of estimated crop acreage.[4] Existing studies that validate small area estimates of non-monetary poverty generated from mobile phone or geospatial data have, by and large, focused on extrapolation, or the use of big data to predict welfare in countries that are not surveyed. This differs from small area estimation, in which the objective is to predict poverty within the area covered by the sample, at a more granular level. To do this efficiently, it is crucial to use the survey data not only to train the prediction model, but also as a direct input into the estimates.

This study therefore makes three main contributions. First, it applies the prevailing framework for small area estimation to combine household survey data with geographically comprehensive geospatial indicators in an efficient way. Second, it evaluates the extent to which incorporating geospatial variables at the subarea level improves the precision of small-area poverty estimates. Finally, it assesses which of the commonly used SAE models – unit-level models (Elbers, Lanjouw, and Lanjouw 2003. Molina and Rao, 2010), and the Fay-Herriot area-level model (Fay and Herriot, 1979) – are best suited for combining survey and geospatial data to produce efficient and accurate estimates of both area-level poverty rates and the uncertainty associated with them.

We test these models in the context or generating small area estimates of non-monetary poverty in two developing countries, Sri Lanka and Tanzania. These countries were selected due to the availability of both census and survey data with subarea-level identifiers and matching polygon shapefiles close to the village level. The welfare prediction model uses survey data to estimate the value of a household welfare index as a function of subarea characteristics, and therefore differs from standard small area estimation models that predict welfare using household characteristics. The resulting estimates provide a large efficiency gain compared with direct survey estimates.

We mainly consider Empirical Best Predictor (EBP) models, which have a long history in small area estimation and can accommodate subarea-level auxiliary data to produce estimates of poverty rates and their mean squared error for target areas. EBP models differ from the "ELL method" (Elbers, Lanjouw, and Lanjouw 2003), that has traditionally been used for poverty mapping by the World Bank because they incorporate additional information from the survey by conditioning the area effect on survey residuals.[5] This can lead to substantial efficiency gains when the variation in welfare or poverty contributed by the sample is large, relative to the variation contributed by the auxiliary data. On the other hand, a drawback of EBP models is the required assumption that the stochastic error terms in the model follow a normal distribution. While this normality assumption can be tested by examining the residuals from the prediction model, violating the assumption will introduce bias into the estimates.

In the traditional case when the auxiliary data are household-level data from a census, the sample is a small fraction of the size of the census and EBP models typically lead to minor efficiency gains.[6] In these cases, it is not obvious that EBP models are preferred to simulations based on standard random effect models that do not condition the area effect on the sample but allow for departures from normality. When the auxiliary data are only available at the subarea level, however, the effective size and statistical power of the auxiliary data are much less than that of a full household census. The share of variation contributed by the sample therefore becomes larger, and as shown in results presented below, EBP

[4] Battese, Harter, and Fuller (1988), Erciulescu et al (2019).
[5] For example, the methods described in Elbers Lanjouw and Lanjouw (2003) and Tarozzi and Deaton (2009) are not EBP models.
[6] Haslett (2016).

models provide a large gain in efficiency compared with traditional ELL methods that do not incorporate Empirical Best methods.

We compare the estimates generated by a household EBP model with direct estimates obtained solely from the survey, as well as the well-known Fay-Herriot area-level model.[7] The latter sacrifices precision by discarding the variation in the geospatial indicators across subareas within target areas. Once the area estimates are obtained from the household EBP and Fay-Herriot models, we compare them to non-monetary poverty rates calculated directly from the full census. The availability of census data provides a credible benchmark for assessing how different methods and their variants perform, both in terms of the accuracy of both the small area point estimates and their confidence intervals. We compare the predictions from different methods in terms of their precision, their accuracy, and the extent to which the estimated confidence intervals contain the true value, which is known as the coverage rate.

The main result is that incorporating remote sensing data in an EBP framework substantially improves the accuracy and precision of small area estimates of non-monetary poverty, largely by incorporating information from non-sampled subareas.[8] This comes at no cost to coverage rates in Sri Lanka and a moderate cost in Tanzania, compared with standard direct estimates. The corresponding efficiency improvement is comparable to roughly tripling the size of the survey in Sri Lanka and quintupling it in Tanzania.[9] The small area estimates moderately underestimate mean squared error by, first, failing to account for uncertainty in estimated variance parameters from the model, second, by incorrectly assuming that sample observations are independent within areas, and third, by failing to incorporate sample weights in model estimation. Estimated coverage rates remain respectable, however, at 75 percent in Tanzania and 84 percent in Sri Lanka when benchmarked. This is comparable to the 76 percent coverage rate in both countries when using standard direct estimates. In Tanzania, the estimates from the unit level model are roughly as accurate and moderately more efficient than those from the area-level Fay-Herriot model. In Sri Lanka, where the poverty rate is low and the Fay-Herriot model is less predictive, the estimates from the unit-level model are substantially more accurate and efficient than the Fay-Herriot estimates.

These results are robust to a variety of robustness checks that explore alternative implementation options, including the absence of benchmarking and the use of a noisier welfare measure. When using the noisier welfare measure, the gain in efficiency is not as large in Sri Lanka, on the order of doubling the size of the sample, but the predictions remain accurate and coverage rates remain high. Finally, to further probe the robustness of the results, we implement both a model-based simulation and a design-based simulation of 250 randomly selected samples covering three regions in Northeast Tanzania. The small area estimates perform exceedingly well, while the design-based simulation also shows that the small area estimates substantially improve on the direct estimates in terms of efficiency and accuracy while maintaining high coverage rates.

[7] See Fay and Herriot (1979).

[8] While the main efficiency improvements occur by incorporating information from non-sampled sub-areas, there are also minor efficiency improvements from combining sample data with synthetic predictions in sampled sub-areas.

[9] This is based on a comparison of the average coefficient of variation between the standard direct estimates and the household level model reported in Table 10.

The reminder of the paper is organized as follows. Section II describes the data. Section III describes the methodology and estimators that are evaluated. Section IV assesses results in terms of efficiency, accuracy, and coverage. Section V considers a variety of robustness checks of the main method. Section VI considers the results of model-based and design-based simulations, and section VII concludes.

II. Description of the data

A. Constructing measures of non-monetary welfare and poverty in the census

The analysis is conducted using measures of non-monetary poverty constructed from the Sri Lankan and Tanzanian census, which were each conducted in 2012. Non-monetary poverty, unlike monetary poverty, is directly observed in the census population and can therefore provide a benchmark for evaluation. For each country, we identified a set of household asset and demographic proxies for welfare. Principal component analysis, weighted according to household size, was used to derive a score for each proxy welfare indicator. The welfare proxies used in the index and their estimated scores, or loading factors, are listed in Table 1. Each household's non-monetary welfare measure is obtained by summing the product of each loading factor and the household's value of the associated variable. No subsequent adjustment for household size was made, reflecting the non-rival nature of the index variables within the household, although household size is included as a welfare proxy in the index for both countries. The results are broadly reasonable, as higher levels of education for the household head are associated with higher non-monetary welfare in both counties, assets and dwelling conditions are positively associated with welfare in Sri Lanka, and non-agricultural work is strongly associated with non-monetary welfare in Tanzania.

This measure of non-monetary welfare can be compared with a poverty line threshold to classify each household as poor or non-poor. We set the threshold at the 4th percentile of non-monetary welfare in Sri Lanka and the 20th percentile in Tanzania, to reflect prevailing national monetary poverty rates. In other words, households were classified as non-monetarily poor if their non-monetary welfare, defined as the score of the first principal component, fell in the bottom four percentiles in Sri Lanka and the bottom quintile in Tanzania. Below in the robustness checks section, we also consider results when the poverty rates of the two countries are swapped.

Table 1: Factor loadings for census-based welfare index by country

Sri Lanka		Tanzania	
Variable	Scoring Coefficients	Variable	Scoring Coefficients
Household size	0.09	Head Literate	0.45
Dependency ratio		Head Ever attended school	0.45
Children 0-14 and 65+	0.04	Head age	-0.20
Children 0-14 Only	-0.05	Household size	-0.22
Gender ratio	0.04	Dependency ratio	-0.41
Household education		Male head	0.12

No schooling	-0.07	Non-agricultural work	0.29
Up to grade 5	-0.14	Non-livestock work	0.25
Grade 5-10	-0.31	Non-fishing work	0.01
O or A level	0.27	No disability	0.07
College Degree or higher	0.16	Cash transfer beneficiary	0.13
Household Assets			
House	0.04		
Computer	0.27		
Landline phone	0.23		
TV	0.28		
Housing characteristics			
Roof	0.14		
Private Toilet	0.18		
Wall	0.19		
Waste disposal	0.15		
Safe water	0.13		
Main cooking fuel is wood	-0.23		
Electricity for light	0.27		
Head Education			
No schooling	-0.11		
Up to grade 5	-0.17		
Grade 5-10	-0.13		
O or A level	0.29		
College Degree or higher	0.13		
Age of head in years	0.04		
Head employment status			
Unemployed	-0.0005		
Public sector	0.15		
Private sector	-0.6		
Out of labor force	-0.04		
Male head	0.11		
Head marital status			
Unmarried	-0.04		
Married	0.12		
Widowed	-0.10		
Divorced	-0.06		
Number of variables	35		11

Notes: Table contains factor loadings from the first principal component, estimated in the census weighting by household size. Dependency ratio is equal to the ratio of non-prime age adults (0 to 14 and 65+) to household size.

B. Constructing synthetic household surveys

With a measure of non-monetary welfare and poverty for each census household in hand, we turn to drawing a synthetic household survey from each country's census. The synthetic survey, along with the auxiliary geospatial data, are key inputs into the small area estimation procedures. To draw the synthetic survey, we utilize the actual two-stage sample conducted by the National Statistics Offices for two household budget surveys: The 2018 Tanzania Household Budget Survey, and the 2016 Sri Lanka Household income and Expenditure Survey. These surveys were merged with the census at the subarea level, which is the GN Division in Sri Lanka and the village in Tanzania. After retaining the GN Divisions and EAs present in the budget survey, we randomly select census households in each matching EA to match the number of households in each EA for each survey. Finally, we merged the sample weights from the household budget surveys for each subarea. Essentially, this procedure draws a survey that mimics as much as possible the sample drawn by the NSO for the budget surveys.

C. Remote sensing data

The auxiliary data for the small area estimation exercise are drawn from a large candidate pool of satellite-based information, most of which is derived from publicly available layers and imagery. These include night-time lights from the Visible Infrared Imaging Remote Sensor (VIIRS), at a spatial resolution of 15 arc-seconds, precipitation data from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS), elevation and slope taken from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) satellite, global forest cover change from Hansen (2013) and estimates of built-up area from the Global Human Settlement Layer (GHSL). From this last layer, 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.

The Sri Lanka indicators were also supplemented by a variety of spatial "texture" features 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 and is the highest spatial resolution (10m per pixel) optical data that are publicly available. The "texture" features were calculated using Jordan Graesser's Sp.Feas package, an open-source Python library for processing contextual image features from satellite imagery. These contextual features have been shown to be strongly correlated with poverty and population density (Engstrom et al, 2017 and 2019a). The following contextual features were considered as candidate predictors:

- **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, which 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. For most features, we use scales of 3, 5, 7, which correspond to squares of 3 pixels by 3 pixels, 5 pixels by 5 pixels, and 7 pixels by 7 pixels.

The remote sensing data used for Tanzania, while largely derived from publicly available imagery, was supplemented by proprietary data on building footprints produced by Ecopia and Maxar.[10] The publicly available layers considered for Tanzania included:

- **Night-time lights**, taken from the same VIIRS imagery used for Sri Lanka
- **Population estimates** taken from WorldPop, which is in turn derived from the 2011 census and distributed using measures of built-up area.
- **Precipitation estimates** taken from Weilcott and Matsuura (2018)
- **Elevation and Slope** taken from Jarvis (2008) and downloaded from AidData's geoquery database
- **Built-up area** from the same GHSL data used for Sri Lanka, for the same periods.
- **Climactic region,** obtained from Kottek et al (2006). This indicator divides land into five main climate groups (tropical, dry, temperate, continental, and polar), with each group being divided based on seasonal precipitation and temperature patterns, based on a raster version of the map available at 0.5° resolution.
- **Crop yield estimates** for Maize, Sorghum, and Rice, from the HarvestChoice Dataverse (Wood-Sichra et al. 2016) capture the estimated yield of various crops (ton/km) at 0.5° resolution.

The full set of satellite indicators used for each country are listed below in Table 2.

Table 2: Indicators and sources for auxiliary satellite data

Satellite indicator	Sri Lanka	Tanzania
Urbanization		
Night-time-lights	VIIRS	VIIRS*
Building footprints		Ecopia and Maxar, via Gates Foundation*
Built-up area		Global Urban Footprint*
Built-up area	Global Human Settlement Layer*	Global Human Settlement Layer
Population		WorldPop
Agglomeration index		Belghith et al (2020) (See Appendix A)
Spatial features	Sentinel 2*	

[10] The team also calculated spatial features using the Sp.Feas package for Tanzania based on Sentinel 2 imagery, but these were found to be highly colinear with the building footprint data and were therefore discarded.

Agro-climactic		
Precipitation	CHIRPS*	Wilmott and Matsuura (2018)
Elevation and slope	ASTER sensor	Jarvis (2008)*
Global forest cover slope	Hansen (2013)	
Climactic region		Kottek et al (2006)
Crop yield estimates		IFPRI Harvest-Choice (Wood-Sichra et al, 2016)
Normalized Difference Vegetation Index	Sentinel 2	Sentinel 2*
Market access		Belghith et al (2020)
Natural Disaster Risk		UNEP/DEWA/GRID-Europe

Note: * indicates that at least one variable from this set was selected in the model of welfare

D. Geographic structure of Sri Lanka and Tanzania

Before describing the methodology used in the study, we briefly review the geographic structure of the two countries. Table 3 shows the subareas, areas, and super-areas for each country. Subareas are the lowest level for which shapefiles were obtained, and are therefore the most disaggregated level for which geospatial data could be linked with the survey. Areas are the target domains for the small area estimates, which are at a higher level than the subareas, but below the level at which the survey is considered to be representative. Finally, super-areas are the lowest level for which the household budget survey is considered to be representative.

Table 3: Geographic structure of Sri Lanka and Tanzania

	Sri Lanka	Tanzania
Sub-areas		
Name	GN Division	Village
Number in country	13,978	16,438
Areas		
Name	DS Division	District
Number in country	331	159
Super-areas		
Name	District	Region
Number in country	25	25

Source: Authors' calculations from 2012 census data.

III. Methods

This section describes the three methods that are evaluated: Direct survey estimates, the Fay-Herriot area-level model, and the household-level model. It also discusses important methodological considerations involved when estimating the household-level model, such as differences between software packages, transforming the dependent variable, the model selection procedure, and benchmarking. The final subsection considers the criteria used to evaluate the performance of different estimators.

A. Direct survey estimates

Direct estimates obtained from the survey serve both as a benchmark against which to assess the increase in precision from incorporating satellite data, and as an input into the Fay-Herriot area-level models. We use two methods to obtain direct survey-based estimates**.** For both methods, the mean poverty estimate for each area is the weighted average, across households, of the dummy variable indicating whether each household is poor. Each household is weighted according to corresponding population weight in the survey. The two methods used to obtain direct estimates differ in how they estimate the variance of the means. The typical method of obtaining variance estimates clusters the residuals by enumeration area, which accounts for positive correlation within enumeration areas but assumes that disturbances across enumeration areas are independent. This assumption underestimates the variance of area-level poverty estimates, particularly if welfare is highly correlated across enumeration areas within target areas.

The Horvitz-Thompson (H-T) approximation offers an alternative, more conservative, approach to estimate the variance of the area poverty rates (Horvitz and Thompson, 1952). This approach is more conservative because it relaxes the standard assumption that households' poverty statuses are independent if they live in different enumeration areas within the same target area. This comes, however, at the cost of imposing an alternative assumption, which is that the probability of one household appearing in the sample is independent of the probability of any other household appearing in the sample. This assumption is clearly violated in standard two-stage sample designs, as the probability of a household appearing in the sample is greater if a household in its same enumeration area also appears in the sample. Nonetheless, the violation of this second assumption underestimates variance by less than the standard assumption of clustered standard errors, and therefore yields less biased variance estimates than the standard approach. We include results for direct estimates based on both the Horvitz-Thompson approach and the standard approach and in results presented below in Table 9, find substantially higher coverage rates for the former.

B. Fay-Herriot area-level model

The Fay-Herriot model was first introduced by Fay and Herriot (1979) to model incomes for small areas of fewer than 1,000 persons in the United States, and is perhaps the best-known and widely used small area estimator. It was derived as an application of the James-Stein (1961) estimator using data aggregated to the target area level, and is also the Empirical Best Linear Unbiased Predictor (EBLUP) of

the area level regression. The small area estimates are a weighted average of the synthetic regression prediction and the direct survey estimate, with the weights inversely related to the estimated variance of the direct and synthetic estimates. Many variants of the Fay-Herriot estimator have been developed that employ transformations and account for spatial and intertemporal correlation, among other refinements.

We estimate the following basic area-level Fay-Herriot model:

(1) $\hat{\theta}_a = X_a\beta + u_a + e_a$

Where $\hat{\theta}_a$ is the direct estimate of headcount poverty in area a obtained from the survey.

X_a is a vector of area-level aggregate remote sensing variables, created by taking population-weighted average of subarea level aggregates.

u_a is a random effect, assumed to be distributed normally with mean 0 and variance σ^2_{ua}

e_a is the sampling error, which is also assumed to be distributed normal with mean 0 and variance σ^2_{ea} . The variance σ^2_{ea} is estimated using the estimates of the variance obtained from the sample as described above.

The variance of the direct survey estimate is estimated as the variance of the direct estimate, calculated using the Horvitz-Thompson approximation, as recommended by Halbmeier et al (2019).

As noted above, the predictions resulting from the Fay-Herriot model can be expressed as a weighted average of the direct estimates and the model predictions.

(2) $\hat{\theta}_a^{FH} = \hat{\gamma}_a\hat{\theta}_a + (1 - \hat{\gamma}_a)X_a\hat{\beta}$

Where $\hat{\gamma}_a = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_{ea}^2)$, or the shrinkage factor. The shrinkage factor is the weight given in each area to the direct estimator relative to the model prediction. This declines as the variance of the direct sample estimate $\hat{\sigma}_{ea}^2$ increases, because the direct sample estimates is less precise. Conversely, the shrinkage factor increases as the estimated random effect becomes less precise and $\hat{\sigma}_u^2$ increases, giving greater weight to the direct sample estimate.

The Fay-Herriot estimator, unfortunately, has several weaknesses for the purposes of generating small area estimates of poverty using spatial auxiliary data. The main shortcoming is that the Fay-Herriot model requires aggregating the auxiliary data to the area level, which discards variation in the auxiliary data at the sub-area level that would increase the precision of the estimates. Implementing a sub-area model that appropriately adapts the Fay-Herriot model to address this issue, such as the one proposed by Torabi and Rao (2014), can incorporate auxiliary data at the sub-area level but is currently constrained by the lack of available software. In addition, the Fay-Herriot model is based on direct estimate of poverty rates, and therefore discards information on the variation in welfare within the poor and non-poor portions of the welfare distribution. Another issue with the Fay-Herriot model is that, when no transformation is applied, the poverty rate is assumed to be a linear function of the predictors. The Fay-Herriot model also faces challenges when no households are poor in a significant number of

areas. For these areas, either the prediction from the auxiliary data must be ignored, the sample must be ignored, or the area must be dropped from the estimation. Finally, the Fay-Herriot model does not account for uncertainty in the survey-based estimates of variance, which tends to underestimate the variance of the prediction.

Despite these limitations, the Fay-Herriot model is much simpler to apply and explain than a unit level model. It generally provides credible estimates and is a standard workhorse of small area estimation. We estimate a Fay-Herriot model by aggregating all indicators to the area level, weighting by population. The Fay-Herriot models are estimated using the Stata FHSAE command (Corral, et al, 2018). We utilize the "Chandra method", which estimates the prediction model using GLS (Chandra et al, 2013).

C. Household-level model

In contrast to area level models, unit level models are specified at the level of the individual unit, which in this case is the household.[11] Although the auxiliary data only varies at the sub-area level, predicting welfare at the household level fully utilizes the information in the sample survey on the level and variability of household welfare. In addition, unlike sub-area models, well-established software packages are available that utilize empirical best methods to estimate unit-level models with area-level mixed effects.

The EBP model can be written as follows:[12]

$$(3) \quad \mathrm{G}(\mathrm{Y}_i) = \beta_1 X_{\mathrm{sa}} + \beta_2 X_{\mathrm{a}} + \beta_3 \mathrm{X}_{\mathrm{r}} + \eta_{\mathrm{a}} + \varepsilon_i$$

Where:

Y_i is the non-monetary welfare for household i, which in each country is the principal component score constructed using the loading factors in Table 1.

$G(Y_i)$ is a transformed version of welfare. Traditionally, ELL uses the log function to transform welfare, but any monotonic transformation can be used. As described in greater detail below, we use an ordered quantile normal transformation for the primary set of results.

X_{sa} is a set of satellite-derived indicators that vary at the sub-area (village or GN Division) level.

X_a is a set of satellite-derived indicators that vary at the area level. The areas sub-districts in Sri Lanka and districts in Tanzania. These predictors are important to include to reduce the variance of the area random effect and thereby mitigate bias in the estimated mean squared error (Elbers, Lanjouw, and Leite, 2008).

X_r is a set of regional dummy variables indicating super-areas for which the survey is considered to be representative. The Sri Lanka specification also includes dummies for the rural and estate sectors, while the Tanzania specification includes a rural dummy.[13]

[11] See Morris (1983) for a comprehensive review of EBLUP models, and Jiang and Lahiri (2006) for a comprehensive review of their application to small area estimation.
[12] See Molina and Rao (2010) and Tzavidis et al (2018).
[13] Sri Lanka is unique in that the national statistics office designates certain areas as the "Estate Sector". The estate sector mainly consists of tea plantations and is historically economically disadvantaged.

η_a is a mixed effect specified at the area level, assumed to be distributed normally. Although traditionally ELL has specified the area effect at the cluster or enumeration area level, doing so underestimates the mean squared error by failing to account for correlation in model error across clusters within areas (Marhuendra et al, 2018, Das and Chambers 2018).[14] η_a is conditioned on the sample values of observed welfare, denoted y_s.

ε_i is a household level idiosyncratic error term, also assumed to be distributed normally.

A key feature of the EBP framework is that the random area effect η_a is conditioned on the sample values of observed welfare y_s. This, along with the assumption that $\sigma^2\eta$ is normal, allows for the use of the following model for generating simulated welfare:

(4) $G(Y_i^*) = \beta_1 X_{sa} + \beta_2 X_a + \beta_3 X_r + \tilde{\eta}_a + \eta_a^* + \varepsilon_i$

(5) $\tilde{\eta}_a = E[\eta_a | y_s]$

(6) $\eta_a^* \sim N\left(0, \sigma^2\eta(1-\gamma_a)\right)$

where

(7) $\gamma_a = \frac{\sigma^2\eta}{\sigma^2\eta + \sigma^2\varepsilon/N_a}$.

γ_a is the shrinkage factor due to conditioning the random effect on the sample mean for area a. N_a is the number of sample observations for area a. The shrinkage in variance due to conditioning the random effect on the sample will be higher when N_a is large or when the variance of the residual ε is low, because in these cases the sample contributes more information about the unexplained component of welfare. Shrinkage falls as the estimated within-area correlation in welfare $\sigma^2\eta$, declines, because in those cases a larger share of unobserved welfare is idiosyncratic to households rather than common to areas, so the sample residuals convey less information about the true area mean. Similarly, the shrinkage factor γ declines as the area sample size N_a declines, because the sample contributes less information about the welfare in the area.

This basic structure is appealing in part because it is similar to the Elbers, Lanjouw, and Lanjouw (2003) model that has been used extensively for survey-census poverty mapping for decades, with the main difference being the conditioning of the random effect on the mean sample residuals. The concept of shrinkage in the nested error EBP framework is well-established in the small area estimation context (Battese, Harter, and Fuller, 1988, Jiang and Lahiri, 2006, Molina and Rao, 2010, Van der Weide, 2014). In addition, unlike Bayesian spatial autocorrelation models, nested error random effect models do not impose additional smoothness in model errors across neighboring areas. Spatial correlation is instead modeled entirely through the area random effects, which are conditioned on the sample residuals for that area.

As noted above, it is not clear that Empirical Best methods are preferred in the traditional poverty mapping setting where a household census serves as the auxiliary data. While EB methods benefit from incorporating information from the sample, the sample is typically a small share of the census

[14] Omitting the cluster effect also can lead to bias by simultaneously overestimating the variance of epsilon and the variance of eta, but the resulting biases act in opposite directions and therefore partially cancel each other out.

population, leading to limited benefits from using EB. Meanwhile, an important drawback of EB is the required assumptions that the components of the error term are distributed normally, which can reduce the accuracy of the estimates if the underlying distribution of the residuals is not normal. The choice of whether to use EB in a typical case, when the auxiliary data are a household census, must weigh the cost of assuming normality against the benefit of obtaining additional information from a relatively small sample.

The calculus changes, however, when the auxiliary data are specified at an aggregate geographic level, such as a subarea or area. In this case, EB estimation is essential to effectively combine the survey data with census-based predictions. Intuitively, the auxiliary data contribute less variation when they only vary across sub-areas, compared to the sample, because the sample, unlike the auxiliary data, contains household level information. In the EBP framework described above, the use of predictors aggregated to the sub-area level increases the variance of the area effect $\sigma^2\eta$, which also represents the covariance of unexplained welfare across households within the same area. This covariance rises when using sub-area predictors because all households within each sub-area share the same set of characteristics used to predict welfare in the model. The increase in $\sigma^2\eta$ due to the use of regionally aggregated predictors in turn increases the shrinkage factor γ, leading to a greater increase in precision when using an EBP model versus the traditional ELL estimator.

The assumption of normality is critical in the EBP framework, as a distributional assumption is required to condition the area level random effect on the average residual of the regression in that area. In the EBP model, this conditioning is equivalent in a Bayesian framework to specifying the average of the residuals as a prior distribution for the random effect under the assumption of normality. The normality assumption is also necessary for the Monte-Carlo simulations of the parameters, which are required because headcount poverty is a non-linear function of welfare. In general, software packages generate small area poverty estimates by repeatedly drawing area effects and the household idiosyncratic error from a normal distribution and aggregating the simulation results. This results in biased estimates of poverty rates if the true error terms implied by the model are not in fact normal. Below, we discuss a monotonic transformation to the welfare measure that makes the assumptions that the area effect and the household error term are distributed normally more palatable.

D. Software packages to estimate EBP models

Table 4 summarizes the three software packages that we utilize to estimate household level models. The primary set of results reported below are estimated using a modified version of the R EMDI package (Kreutzmann et al, 2018).[15] We use this package for the main results because it is well-documented, flexible, and fast when running on powerful computers that can exploit parallel processing. In addition, it appends the sample to the census in order to generate slightly more accurate estimates. This package, like others designed for poverty mapping, estimates the mixed model (3) and then repeatedly draws both area effects and household error terms to simulate welfare. Simulated welfare is compared to the

[15] The R EMDI package is available on CRAN. It is an extension of the R SAE package, and runs substantially faster because it allows the user to parallelize the bootstrap simulations across several cores.

poverty line to determine the poverty status of each household in each simulation, which is aggregated across simulations to estimate poverty in each area. Estimates of mean squared error are obtained through a parametric bootstrapping approach, which is described in detail in Gonzalez-Manteiga, et al (2008) and Molina and Marhuenda (2015). We leverage the open nature of the code to make one modification to the EMDI package related to the use of weights, which allows for household size to be used as a weight when aggregating the simulated household welfare outcomes into area-level poverty estimates.[16]

As a robustness check, we also utilized the second and third versions of an alternative software package, the Stata SAE package (Nguyen et al, 2018).[17] The main differences between the packages are summarized in Table 4, and discussed in Annex A. These differences in design and methodology across different packages, particularly the choice of a non-parametric versus parametric bootstrap method, affects the resulting estimates of poverty rates and their mean squared errors. Comparing simulated small area estimates against actual census data provide evidence on how these differences affect estimated poverty rates and mean squared error in two use-case scenarios.

Table 4: Differences between Stata SAE and R SAE package

	Stata ELL		Modified R EMDI
	Version 2	Version 3	
Heteroscedasticity correction	Optional	Optional	Not available
Type of bootstrap	Traditional Non-Parametric clustered bootstrap with varying sample composition.	Parametric bootstrap with constant sample composition.	Parametric bootstrap with constant sample composition.
Model fitting method	Henderson method 3 with non-parametric bootstrap	Henderson method 3	Maximum likelihood estimation
Empirical Best estimation	Optional	Required	Required
Sample weights	Optional	Optional	Not available
Sample appended to census for simulations	Not available	Not available with sample weights	Required

E. Normalizing transformation

Welfare typically follows a right-skewed distribution, and it is therefore standard practice to transform the welfare indicator before implementing small area estimation models based on simulated welfare. The most common transformation is to take the logarithm of welfare, following Elbers, Lanjouw, and

[16] This modification involved taking a weighted mean instead of a simple mean across areas when aggregating households' simulated poverty status for each area. Code is available upon request from the authors.
[17] Version 2 was released in September 2019 and is available for download at: https://github.com/jpazvd/sae. Version 1 is available through the Stata SSC archive. Version two implemented a clustered bootstrap at the level of the enumeration area, using the traditional non-parametric bootstrap. Version 3 is a comprehensive update that implements a parametric bootstrap approach similar to the R EMDI package (Corral, et al, 2020).

Lanjouw (2003). A recent review article considers the issue of transformations for small area estimation in detail (Tzavidis et al, 2018). It reports design-based simulation results using Mexican data that considered alternative transformations, including log-shift and Box-Cox transformations (Box and Cox, 1964). A log-shift transformation involves adding a constant to welfare prior to taking the log, while the Box-cox transformation can be written as: $\frac{y^{\lambda}-1}{\lambda}$ for $\lambda \neq 0$, and log y for $\lambda=0$. Tzavidis et al (2018) conclude on the basis of the simulation results that log-shift transformations tend to perform well. Stata has implemented two commands that can quickly estimate the parameters for the log shift and Box-Cox transformations that minimize the skewness of the welfare distribution. These commands, however, only go part of the way towards normalization, because they cannot guarantee that the resulting distribution will have a kurtosis of three, which is characteristic of a normal distribution. Furthermore, the log shift transformations is not monotonic when the underlying distribution is left-skewed.[18] Finding an appropriate monotonic transformation for the welfare variable is important, as failure to satisfy the normality assumption can lead to systematic bias in EBP models. While benchmarking can be used to at least partially correct this bias, relying on benchmarking becomes more difficult to justify as the extent of the bias increases.

For these reasons, our preferred estimates utilize a monotonic transformation called the "ordered quantile normalization" to transform welfare. This method transforms the scaled rank of the welfare variable to conform with a normal distribution. Variants of this method date back more than 70 years (Bartlett, 1947, Van der Waerden, 1952). Among several methods proposed in the literature, the ordered quantile normalization most consistently transforms an underlying variable to follow a normal distribution (Peterson and Cavanough, 2019). We utilize the Ordernorm function in the Bestnormalize R package to implement this transformation. The method produces an exactly normal distribution of the transformed variable when the original welfare measure contains no ties, irrespective of its underlying distribution. While a normally distributed outcome variable does not guarantee normality of the residuals, it helps make the distribution of residuals closer to normal, and the ordered quantile normalization leads to smaller discrepancies between the official national poverty rate estimated from the survey and the weighted mean of the small area estimates.

To calculate small area estimates of the poverty gap, poverty severity, or inequality, it is critical to retransform the estimates back to the original welfare metric, as is done in the ELL method by exponentiating log welfare in each simulation. However, this retransformation is not necessary when estimating headcount poverty rates. This is because households' poverty status is preserved under any monotonic transformation applied to both the welfare measure and the poverty line threshold. We therefore set the poverty line to the percentile of the transformed welfare measure that corresponds to the assumed national poverty rate, which is 4 percent in Sri Lanka and 20 percent in Tanzania. Because the ordered quantile normalization transformation is based on household ranks within the distribution, applying it entails information loss regarding the shape of the distribution between household ranks, and alters the estimated parameters of the model. To verify that this loss of information is minor compared with the benefits of normalizing the welfare measure, we also consider in the robustness section below a more traditional Box-Cox normalization procedure designed to minimize skewness.

[18] When the underlying distribution is left-skewed, the skew-minimizing transformation becomes $\log(-y+k)$, which is not monotonic.

F. Model selection using the Lasso

Model selection in small area estimation remains an unsettled issue and has traditionally been treated as both art and science. Practice can vary widely. Tzavidis et al (2018), for example, select only 6 covariates present in Mexican census and survey data, based in part on the Akaike Information Criterion (AIC) from a standard OLS model. Most small area applications using ELL, however, use a considerably larger set of variables. Zhao and Lanjouw (2008), for example, note that many successful applications include less than 20 household level variables. However, they also advocate including cluster-level means, which can boost the number of variables significantly past 20 or 30.

In recent years, the least absolute shrinkage and selection operator (LASSO) has become an increasingly popular tool for selecting prediction models. Small area estimation offers a natural application of the "post-lasso" procedure, in which the variables selected via the lasso are used as predictors into the mixed model used to estimate the parameters of the small area estimation model.[19] The post-lasso procedure offers the benefit of a convenient and data-driven approach to model selection that, in its most popular variant, maximizes out-of-sample predictive accuracy, while roughly equalizing in and out of sample R^2 to prevent overfitting the model to the sample. The LASSO framework formalizes and extends the advice in Zhao and Lanjouw (2008) to use out-of-sample cross-validation to ensure that the model is stable and predicts well out of sample. Several varieties of LASSO have been implemented in popular statistical software. We use the plugin lasso estimator, also known as the rigorous lasso, implemented in Stata version 16 (StataCorp, 2019). This variant of lasso is computationally faster and more parsimonious than others that use cross-validation to minimize mean out-of-sample prediction error, and the resulting models maintain strong predictive power.[20]

G. Benchmarking

Benchmarking refers to the procedure of forcing model-dependent estimators to agree with the direct sample estimates for super-areas, defined as the lowest geographic level for which the sample is considered representative. Benchmarking can be desirable to ensure that the population-weighted average of small area poverty estimates, when aggregated to super-areas, match official published statistics derived from the survey. Simulations indicate that in the cross-sectional context, simple "ratio" benchmarking performs equally well as other methods and there is no significant loss of efficiency from benchmarking in general (Pfefferman et al, 2014, Wang et al, 2008). We therefore apply a simple ratio benchmarking procedure by multiplying the estimated headcount rate in each area by a scaling factor unique to each super-area, for both the Fay-Herriot and household level model estimates. The scaling

[19] See Hastie et al (2015) for a review of the lasso and related estimators, and Belloni and Chernozhukov (2013) for more details on the post-lasso procedure.

[20] We use the variant of the Stata plugin estimator that allows for heteroscedastic errors. The plugin method uses an iterative formula to select the lambda shrinkage parameter in the lasso instead of a grid search. See StataCorp (2019), Belloni and Chernozhukov (2011), and appendix A of Belloni, Chernozhukov, and Hansen (2014) for technical details regarding the implementation of the plugin lasso estimator.

factor is defined as the ratio of the estimated poverty rate obtained from the survey to the population-weighted mean poverty rate of the small area estimates, for each super-area.

This leaves open the question of how to adjust the mean squared error estimates while benchmarking. Ideally, to obtain accurate estimates of the mean squared error, benchmarking would be performed by the small area estimation package within each bootstrap replication. Unfortunately, this option is not currently implemented in available software packages. As a second-best solution, we scale the mean squared error by the square of the same scaling factor that is applied to the point estimates, which leaves the coefficient of variation unchanged by the benchmarking process. Because the point estimates are slightly underestimated in Tanzania and Sri Lanka, this procedure increases the mean squared error and improves coverage rates. However, as a robustness check, we also report results for estimates in which only the point estimates are benchmarked, and in which neither the point estimates nor the mean squared errors were benchmarked.

H. Criteria for evaluating estimates

The synthetic sample, auxiliary data, and methodology can be used to generate small area estimates of headcount poverty and their mean squared error. But how exactly does one evaluate the estimates produced by different small area estimation methods? We examine seven summary statistics, averaged across areas. We give each area equal weight in the evaluation to better reflect the goal of obtaining accurate estimates for each target area rather than estimates that can be aggregated to higher levels. The six summary statistics are divided into four buckets as follows:

- **Mean Poverty Rate and Average Relative Bias.** The former represents the simple average of the predicted area headcount rates across areas, prior to benchmarking. For the direct survey estimates, the mean poverty rate is the unweighted average poverty rate across municipalities. The mean of the small area estimates can differ from this unweighted average, due to bias caused by either violations of the normality assumptions or a non-representative sample. Substantial bias would necessitate a greater reliance on benchmarking, which would in turn make the imperfect adjustment for mean squared errors more problematic. Therefore, estimators with mean poverty rates closer to those of the direct estimates are, ceteris paribus, preferred. Meanwhile, the average relative bias equals the average, across areas of the ratio of the difference between the true and estimated poverty rates to the true poverty rate.

$$\text{ARB} = \frac{1}{N_a}\sum_{a=1}^{N_a}\frac{\widehat{\Theta}_a - \Theta_a^*}{\Theta_a^*}$$

  Where N_a is the number of areas, $\widehat{\Theta}_a$ is the estimated poverty rate for area a, and Θ_a^* is the true poverty rate calculated from the census. Like the average poverty rate, it is a measure of the systematic bias in the estimates rather than accuracy. It differs from the average poverty rate, however, by reflecting any systematic bias at the area level after the benchmarking procedure.

- **Average Mean Squared Error (MSE) and Average Coefficient of Variation (CV).** These measure the degree of uncertainty associated with the point estimates. Estimation of the MSE for the

model-based estimates are generated using a parametric bootstrap under the assumption that the model holds, and much remains to be learned about how model misspecification affects MSE estimates. The CV for each area is defined as the ratio of the square root of the estimated mean squared error to the mean estimated poverty rate, a definition sometimes referred to as CV(RMSE). We report the average CV across areas. CVs are an important indicator of uncertainty because in many cases national statistics offices will not publish statistics when CVs exceed a maximum threshold.[21] Comparisons of the average CV between the model-based and direct estimates should be interpreted with caution, because CVs of the direct estimates are not defined for areas with no poor households in the survey, leading the set of areas included in the average to differ for direct estimates.

- **Correlation and Root Mean Squared Error.** These measure the accuracy of the prediction. The correlation represents the simple correlation between the small area estimates and the actual headcount poverty rates calculated from the full census. The Root Mean Square Error (RMSE) is equal to the square root for the average squared difference between the estimates and the actual census poverty rates

$$RMSE = \sqrt{\frac{1}{N_a}\left(\sum_a \left(\widehat{\Theta}_a - \Theta_a^*\right)^2\right)}$$

  The root mean squared error in this context is a function of the predicted poverty rates and is therefore a measure of accuracy. In contrast, the average mean squared error described above is a measure of precision or uncertainty.

- **Average Coverage rate.** The average coverage rate represents the share of areas for which the estimated 95 percent confidence interval contains the actual census poverty rate, and are a standard metric used to assess the performance of estimators in the literature. [22] The upper and lower bounds of the confidence interval are determined by multiplying the square root of the estimated mean squared error by 1.96 and adding and subtracting it from the point estimate.

IV. Main Results

A. Fay-Herriot area model

Table 5 presents the coefficient estimates from the Fay-Herriot area level model, which is used as a reference point for comparison with the direct estimates and the unit level model.[23] The area-level predictor variables were selected using the lasso plugin method, to prevent overfitting the model to the

[21] For example, the US Census Bureau considers serious data quality issues related to sampling error to occur when the estimated coefficients of variation (CV) for the majority of the key estimates are larger than 30 percent.
[22] See, for example, Tarozzi and Deaton (2009), Demombynes et al (2007), Elbers, Lanjouw, and Leite (2008), Pratesi and Salvati (2007), Hidiroglu and Yu (2016), among many others.
[23] In particular, examining the Fay-Herriot model allows us to quantify the potential gains in efficiency and predictive accuracy from including sub-area level predictors in these contexts.

sample. The set of candidate variables included all area-level remote sensing indicators as well as super-area dummies, although no super-area dummies were selected in Tanzania. In each country, the lasso selected a small number of predictor variables, only 3 in Tanzania and 8 in Sri Lanka. The dependent variable is headcount poverty, expressed as a percentage from zero to one hundred. In Sri Lanka, less annual variability in rainfall as well as a positive rainfall shock in the first quarter are associated with lower poverty rates. In Tanzania, districts with larger buildings are much less poor, as are areas with a larger share of built-up area, while more rural districts tend to be poorer. Despite containing only three variables, the Tanzania model explain half of the variation in non-monetary poverty rates. Meanwhile, the eight Sri Lanka variables explain only 26 percent of the variation in headcount poverty across subdistricts. Overall, the results suggest that the Fay-Herriot estimates, despite using predictors that only vary across area, can significantly improve on the efficiency of the direct estimates, particularly in Tanzania where the model fit is especially good.

Table 5: GLS coefficients from Fay-Herriot estimation

Sri Lanka		**Tanzania**	
Mean nighttime lights in 2016	-0.003	Mean area of nearest 25 buildings.	-19.45***
1990 Built-up area (GHSL)	-0.006	Percent built-up area (Global Urban Footprint)	-2.56*
Standard deviation of rainfall	0.006***	Share rural	0.83
Rainfall deviation from historical mean, Q1	-12.16***	Constant	84.70***
Colombo	-4.55**		
Gampaha	-3.80*		
Matara	-1.44		
Ratnapura	-6.58***		
Constant	-7.37**		
R2	0.26	R2	0.50
Number of Observations	328	Number of Observations	159

Notes: Coefficients based on Fay-Herriot estimation of poverty rates using the Chandra method. Predictor variables selected using the lasso plug-in method.

B. Household-level model

Next, we turn to the unit level model, and examine the results of the post-lasso regression used to estimate the model parameters in Table 6. The first striking result is the overall predictive power of the models. The R^2 from the model is 0.32 in Tanzania and 0.25 in Sri Lana, which is notable given that the explanatory variable, non-monetary welfare, is measured at the household level while all the independent variables vary only at the subarea level. A Shapley decomposition indicates that about half of the model's explanatory power in Sri Lanka and about a third in Tanzania are derived from super-area and sector dummies, which by construction cannot explain any variation in non-monetary areas across areas within super-areas. This illustrates one of the limitations of relying solely on R^2 as a measure of the ability of the model to discern poor areas within the super-areas for which credible survey estimates are already available. Below, we consider an alternative, more informative metric, which is the extent to which the small area estimation procedure increases the precision of area estimates relative to direct survey-based estimates.

At the subarea level, the satellite indicators explain an impressive amount of the variation in average non-monetary welfare. In sub-area level regressions, predictor variables explain 73 percent of the variation in average non-monetary welfare in Tanzania and 59 percent in Sri Lanka. The subarea model R^2 of 0.73 percent for Tanzania is substantially higher than the estimates presented in Jean et al (2016), which find that features derived from satellite data explain 58 percent of the DHS asset index in Tanzania. The stronger predictive power of the model for Tanzania likely results from the use of a richer welfare index as the dependent variable, as well as the use of interpretable features derived from imagery such as building footprints, rather than features optimized to predict night-time lights.[24] Meanwhile, the R^2 of 59 percent for Sri Lanka is remarkably similar to estimates of the explanatory power of monetary poverty reported in Engstrom et al (2017).[25]

Table 6: Post-lasso model of normalized household per capita consumption

Sri Lanka		**Tanzania**	
Variable	Coefficient	Variable	Coefficient
Sub-area variables		**Sub-area variables**	
1990 built-up area	-0.30	Sum of nighttime lights	0.03*
2014 built-up area	1.82***	Mean population (GHSL)	0.01
Standard deviation of rain	0.00	Minimum agglomeration index	-0.01*
Rain Z score, Q2	0.05*	Mean std. dev. of size of 5 nearest buildings	-0.01
Rain Z-score squared, Q4	-0.45	Sum of mean size of 5 nearest buildings	0.30***
Fourier Transform mean, scale 7	0.08***	Number of buildings within 100 m	0.05**
Line Support Region mean, Scale 7	0.47**	% area never built-up, 1975-2015 (GHSL)	-0.25
		% of areas built up 1990 to 2000 (GHSL)	0.13
Area variables		% of areas built up 1975 to 1990 (GHSL)	1.47***
1990 built-up area	2.47***	% of area built-up (GUF)	0.04
Standard deviation of rain	1.06***	Precipitation in 2014	0.00
Rain Z score squared, Q3	-0.49	% Humid tropical rainforest (Kloppen classification)	5.49***
Rain Z-score squared, Q4	-0.46	Standard deviation of NDVI	56.67***
Gabor standard deviation, scale 5	-0.83***		
Histogram of Ordered Gradients standard deviation, Scale 5	-1.74***	**Area variables**	
Line Support Region mean, Scale 7	-0.69**	Mean of nighttime lights	0.15
Structural feature sets mean, scale 7	1.00***	Minimum of nighttime lights	0.48
		Maximum economist costs of drought	0.00

[24] Ayush et al (2020) also finds that the use of interpretable features instead of features derived from transfer learning increased predictive power by 31 percent in Uganda, which is consistent with the stronger performance seen in these models.

[25] Engstrom et al (2017) averages a model-based estimate of poverty derived from the census over large numbers of census households in each village, which makes it similar in practice to a non-monetary measure of poverty.

Super-area variables		**Super-area variables**	
Colombo	2.47***	Morogoro	0.18**
Kalutara	1.06***	Simiyu	-0.21**
Hambantota	-0.49		
Jaffna	-0.46		
Mannar	-0.83***		
Mullaitivu	-1.74***		
Kilinochchi	-0.69**		
Batticaloa	1.00***		
Trincomalee	-0.14		
Puttalam	7.33***		
Anuradhapura	2.47***		
Badulla	1.06***		
Moneragala	-0.49		
Rathnapura	-0.46		
Sector		Sector	-0.34***
Urban	2.65***	Rural	
Estate	-2.02***		
			-1.29***
Constant	-4.49***	Constant	
R^2	0.251	R^2	0.321
Of which: sub-area variables	0.060	Of which: sub-area variables	0.134
area variables	0.058	area variables	0.059
Super-area and sector dummies	0.129	Super-area and sector dummies	0.104
Number of observations	19,570	Number of observations	9,393

Notes: Stars indicate statistical significance at 5, 1, and 0.1 percent levels with robust standard errors. Subarea variables are aggregated across GN Divisions in Sri Lanka and villages in Tanzania. Area variables are population-weighted averages of sub-area variables aggregated to the are level. Super-area variables are district dummies in Sri Lanka and regional dummies in Tanzania. Sector dummies are a rural dummy in Tanzania and urban and estate sector dummies in Sri Lanka. See Table 2 for sources of remote sensing indicators.

For Tanzania, the lasso model selected several building footprints and built-up area measures, as well as measures of night-time lights that capture building and population density. Building counts, night- time lights, and built-up area are all positively associated with welfare. Several climactic variables were also selected, reflecting the dependence of rural areas on favorable rainfall patterns. Higher variance in the NDVI vegetation index, reflecting areas that contain a mix of built-up area and green space, is also positively related to welfare. The Sri Lanka model contains measures of built-up area, rainfall, as well as texture features such as the Fourier Transform and Line Support Region Mean at the subarea level, and other features such as the standard deviation of the Gabor filter and Histogram of Ordered Gradients at

the area level. This is consistent with previous research showing that these texture algorithms or contextual features reflect spatial variability in building and road patterns, as well as poverty.[26]

A Shapley decomposition indicates that the higher R^2 in Tanzania, compared with Sri Lanka, can be attributed to more predictive sub-area variables, which predict 13 percent of the variation in Tanzania as opposed to 6 percent in Sri Lanka. This is likely due to two main factors. The first is that the Tanzanian indicators include measures of building footprints at the subarea level. The second is that Tanzania is a less developed setting in which building density and climactic variables, which most of the satellite variables are capturing, are more strongly correlated with non-monetary welfare. [27]

C. Evaluation Results

The previous section demonstrated that remote sensing indicators are predictive of both area-level poverty and household non-monetary welfare. This section turns to evaluating the small area estimates generated by different methods. Before considering the evaluation results, Table 7 reports model diagnostics for the unit-level model. The geospatial indicators in the model predict about 37 and 30 percent of the variation in household non-monetary welfare in Sri Lanka and Tanzania respectively.[28] The conditional R^2, which reflects the gain in model fit from conditioning the mixed effect on the sample, is about 3 percentage points higher than the marginal R^2 in both countries. The distribution of both the area effect and the household error term is roughly normal, reflecting the success of the normalizing transformation. The Wilks-Shapiro test for normality of the area effect fails to reject the null hypothesis of a normally distributed area random effect in both countries. Figure 1, which plot the quantiles of both the area effects and the household residuals against the quantiles of the normal distribution for each country, confirms that the distribution of both components of the error term is approximately normal, although there are negative outliers in the left tail of the area random effect in Tanzania.

Table 8 begins by presenting the (unweighted) mean poverty rates and uncertainty of the small area estimates, as measured by the mean squared error, multiplied by ten thousand, and the mean coefficient of variation. The table presents the results for direct survey estimates for both the standard method and the Horvitz-Thompson estimates, as well as the Fay-Herriot area model and the unit-level model estimated with the modified EMDI package. All are simple unweighted averages across areas.

[26] Engstrom et al. 2019b, Engstrom et al. 2019c, Sandborn and Engstrom 2016.

[27] In the simulated surveys drawn from the census, the rural dummy explains 24 percent of the variation in household non-monetary welfare in Tanzania, while the urban and estate sectors only explain 6 percent in Sri Lanka.

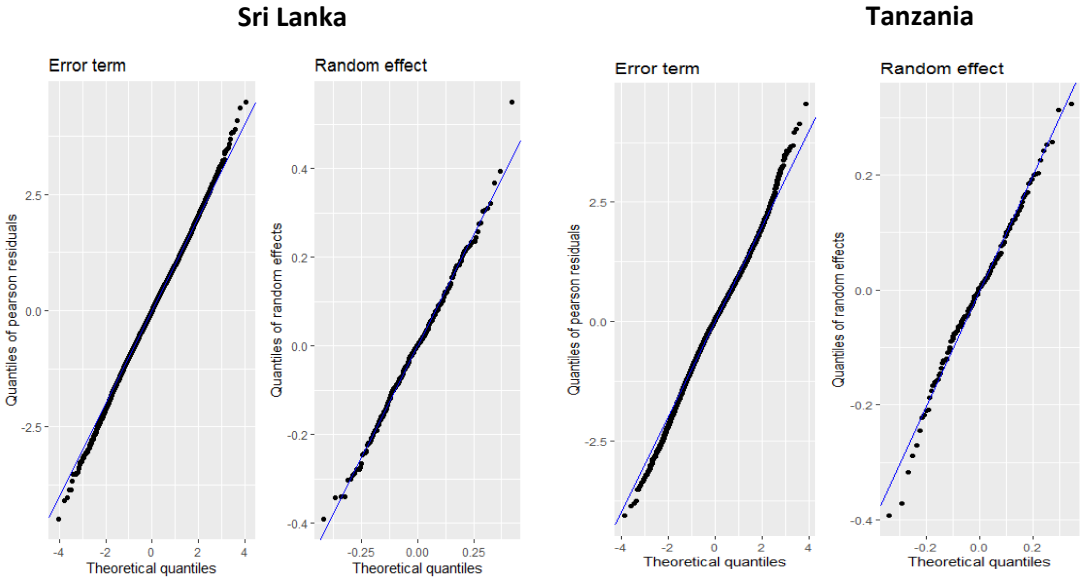
[28] This corresponds to the marginal R^2 in Table 7. The R^2 results differ slightly from the post-lasso model because the small area estimation model does not use weights.

Table 7: Details and Diagnostics for unit level models

	Sri Lanka	Tanzania
Poverty rate	4.0%	20.0%
Number of variables selected for model	28	19
Number of households in sample	19,570	9,393
Number of households out of sample	4,190,436	8,901,279
Number of sub-areas in survey	2487	786
Number of sub-areas in census	13,985	14,981
Number of areas in survey	328	159
Number of areas in census	331	159
Unit-level model diagnostics		
Marginal R^2	0.266	0.299
Conditional R^2	0.296	0.323
Skewness of area effect	0.101	-0.26
Kurtosis of area effect	3.385	3.60
Wilks-Shapiro P-value	0.558	0.316
Skewness of household error	-0.05	-0.12
Kurtosis of household error	3.23	3.51
Variance of estimated area effect	0.036	0.024
Variance of estimated household residual	0.723	0.682
Percentage of variance of error term due to area effect	5.0%	3.5%

Notes: Table reports number of selected variables, number of observations, number of sub-areas and target areas, conditional and marginal R2, indicators of normality for area random effect and household residuals, and the percentage of total variance in the error term accounted for by the area random effect.

Figure 1: Quantile vs. Quantile plots of household error terms and area random effects



Notes: Figures report normal quantile-quantile plots of estimated household welfare residuals (error term) and area random effects (random effect).

The results for the mean poverty predictions in Table 8 show that the non-benchmarked Fay-Herriot estimates underestimate poverty, by a slight amount in Sri Lanka but by a more significant amount in Tanzania. The difference in precision is more striking. The Fay-Herriot estimators are significantly more precise than the Horvitz-Thompson direct estimates in both countries. Relative to these more conservative direct estimates, the Fay-Herriot procedure shrinks the mean square error by nearly half in both countries. However, the household level model, compared with the Horvitz-Thompson estimates, shrinks the mean squared error substantially more, by 70 percent in Sri Lanka and 85 percent in Tanzania.

Table 8: Average Mean Squared Error (MSE) and coefficient of variation (CV) of target area estimates of headcount non-monetary poverty, by country and method

Mean and Uncertainty	Sri Lankan subdistricts			Tanzanian districts		
	Mean poverty	Mean MSE	Mean CV	Mean poverty	Mean MSE	Mean CV
Direct survey estimates						
Horvitz-Thompson approximation	4.0	15.5	65.2	20.0	73.7	43.3
EA-Clustered variance estimates	4.0	10.9	58.6	20.0	59.2	38.1
Area level Fay-Herriot model	3.7	8.3	47.9	16.8	35.8	34.2
Household-level EB model	3.7	4.6	32.0	18.6	11.1	17.5

Notes: Mean poverty refers to the average of area headcount rates prior to benchmarking. Mean MSE is the mean across areas of the mean squared error, estimated using a parametric bootstrap approach, multiplied by 10,000

Table 9 shows the evaluation results against the area poverty rates derived from the census. Average relative bias is highest for the Fay-Herriot estimator in both countries, and lowest for the direct estimates. Of greater interest for evaluating accuracy, however, are the results for correlation and root mean squared error. In both countries, the small area estimates are far more accurate than the direct estimates, but the comparisons between the household level model and Fay-Herriot model is mixed. In Sri Lanka, the household model estimates are more highly correlated with the census than the Fay-Herriot estimates by a moderate amount (0.88 vs 0.84), but slightly less strongly correlated in Tanzania (0.876 vs. 0.883). The Fay-Herriot predictions have a lower root mean squared error in Sri Lanka, reflecting fewer large outliers, but the root mean squared error is virtually equal in Tanzania. Overall, with regards to accuracy, both the household level model and the Fay-Herriot model perform well.

The household model estimates are substantially more precise than the Fay-Herriot model and the direct estimates, but the tighter confidence intervals around the household model estimates may lower coverage rates. The last column of Table 8 shows the coverage rate, which for the household model in Sri Lanka is 84.3%. This is slightly higher than the Horvitz-Thompson estimates (82.2%), moderately higher than the clustered direct estimates (76.1%) but substantially lower than the Fay-Herriot estimates (91.8%). For Tanzania, the coverage rate for the household level model is 74.8%. This is the lowest of the four estimators but only modestly lower than that of the standard enumeration-area clustered estimates (76.1%). The Horvitz-Thompson and Fay-Herriot estimators achieve much higher coverage rates of 85.5% and 91.8%, respectively.

Table 9: Average Relative Bias, Average correlation, Root Mean Squared Error, and Coverage rate of target area estimates of headcount non-monetary poverty, by country and method

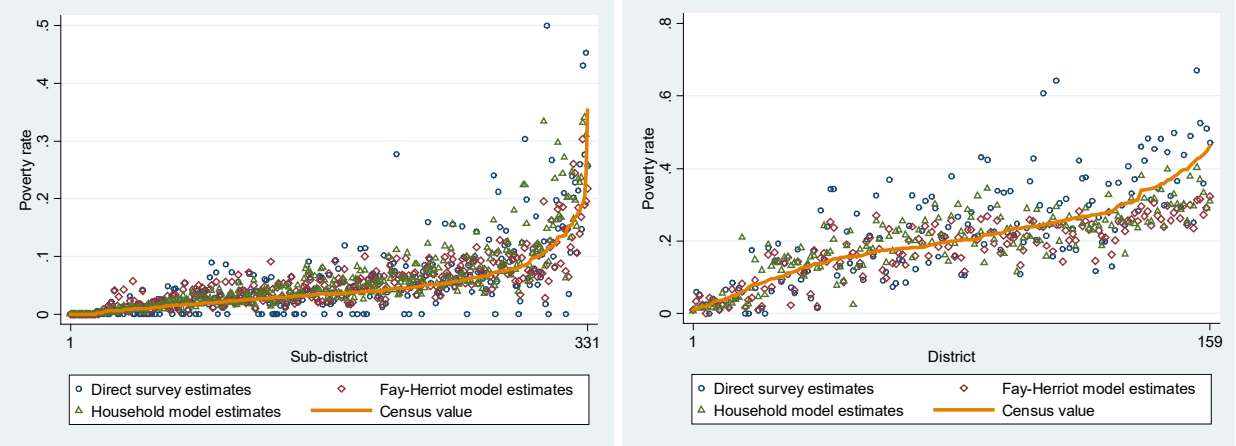
Accuracy and Coverage rate	Sri Lankan subdistricts				Tanzanian districts			
	ARB	Correlation	RMSE	CR	ARB	Correlation	RMSE	CR
Direct survey estimates								
H-T	0.257	72.5	0.049	82.2	0.043	77.0	0.088	85.5
EA-Clustering	0.257	72.5	0.049	76.1	0.043	77.0	0.088	76.1
Area level Fay-Herriot model	0.395	84.1	0.026	91.8	0.071	88.3	0.052	91.8
Household-level EB model	0.293	87.6	0.034	84.3	0.062	87.6	0.053	74.8

Notes: ARB refers to average relative bias, which is the average ratio of the difference between estimated and census poverty rates to the census poverty rate. Correlation refers to the unweighted correlation across areas between estimated and census non-monetary poverty rates. RMSE refers to root mean squared error, which is the square root of the average squared difference between estimated and census poverty rates. CR stands for coverage rate, which is the share of areas for which the estimated 95 percent confidence intervals for the area non-monetary poverty rate contains the census non-monetary poverty rate. Point estimates have been rescaled to match direct survey estimates at the regional level and mean squared error adjusted accordingly. H-T refers to the Horvitz-Thompson approximation, while EA clustering refers to standard variance estimates clustered by enumeration area.

Figures 2 and 3 give a detailed look at the point estimates and mean squared errors for the area poverty estimates for each method. Sri Lanka is shown on the left panel and Tanzania on the right, and in each country the areas are ordered according to their non-monetary poverty rate in the census. The results clearly show that both Fay-Herriot and Household-level models greatly improve on the accuracy of the direct estimates, especially in areas with higher poverty rates. In Sri Lanka, many of the direct estimates are zero, even in high-poverty areas. The comparison between household models and Fay-Herriot models is less clear, but the Fay-Herriot estimates appear to be more prone to overestimate poverty for low-poverty areas and underestimate poverty for high-poverty areas.[29] Figure 3 presents the mean squared errors produced by the different methods. It clearly demonstrates that the household model estimates are substantially more precise than both the Fay-Herriot estimates and the direct estimates in both countries.

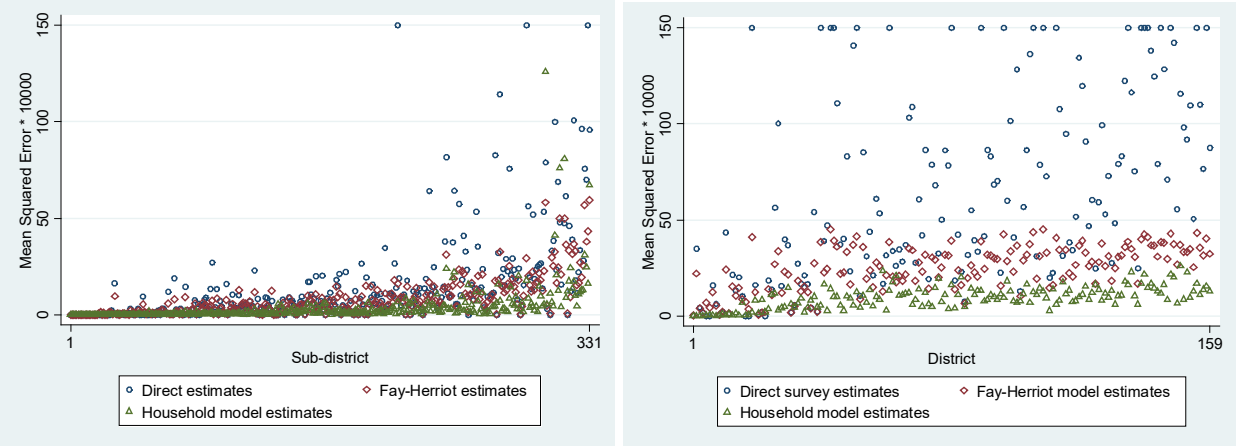
[29] As described above, the Fay-Herriot estimates use the predicted poverty rate from the model in areas where no sampled households are poor.

Figure 2: Comparison of area poverty estimates by method for Sri Lanka (left) and Tanzania (right):



Notes: Figures show predicted asset-based poverty rates generated by household-level model, Fay-Herriot model, and Direct Survey estimates, in comparison with actual asset-based poverty rates calculated from the census

Figure 3: Comparison of estimated Mean Square Error by area and method for Sri Lanka (left) and Tanzania (right)



Notes: Figures show mean squared errors of predicted asset-based poverty rates generated by household-level model, Fay-Herriot model, and Direct Survey estimates. Direct survey estimate MSEs top-coded at 0.015.

The significantly lower mean squared errors generated by the household level model is associated with a moderate reduction in coverage. This begs the question of whether the household level model remains more precise than others after equalizing coverage rates. To shed light on this, we perform a subsequent adjustment. This adjustment artificially inflates the estimates MSEs from each method by multiplying them by a constant scale factor greater than one, with the scale factor manually selected to achieve a 95 percent coverage rates. This occurs after the benchmarking and is therefore an additional adjustment to the mean squared error. Table 10 presents the results. The first column shows the scale factor that was multiplied by the mean squared error to achieve a 95 percent coverage rate, while the

Table 10: Measures of uncertainty when mean squared errors are rescaled to achieve 95 percent coverage rate

	Sri Lankan subdistricts			Tanzanian districts		
	Scale factor	Mean MSE	Mean CV	Scale factor	Mean MSE	Mean CV
Horvitz-Thompson direct estimate with rescaled MSE	N/A	N/A	N/A	2.5	460.3	108.3
Fay-Herriot model with rescaled MSE	2.46	20.4	75.3	1.44	51.5	41.1
Household-level model with rescaled MSE	2.37	10.3	47.9	3.61	40.2	33.2

Notes: Scale factor refers to a constant scaling factor that was multiplied by the estimated mean squared error to inflate the MSEs to achieve 95 percent coverage rates. Mean MSE is multiplied by 10,000.

second and third columns give the post-adjustment average MSE and CV.[30] In Sri Lanka, the household level model is substantially more efficient than the Fay-Herriot model when the mean squared errors are rescaled, as the average mean squared error is 45 percent lower. The difference between the Fay-Herriot and the household level model results is smaller in Tanzania, because the Fay-Herriot model performs better. The household level model remains moderately more efficient after the rescaling, however, as the average mean squared error and average CV are each about 20 percent lower.

Finally, table 11 compares the precision of the unit level model with the direct survey estimates for super-areas, which are districts in Sri Lanka and regions in Tanzania. The super-areas are levels for which the household survey is considered representative and survey-based monetary poverty rates are routinely published. In Sri Lanka, where there are 331 target areas, the mean squared errors for the subdistrict estimates are about sixty percent larger than the direct survey estimates for districts. In Tanzania, meanwhile, there are only 159 areas and the same number as super-areas as Sri Lanka. In this case, the average mean squared error of the small area district estimates is slightly lower than the direct estimates for regions, while the mean CV is modestly higher. This type of comparisons is useful to inform decisions by national statistics offices of whether small area estimates are sufficiently precise to publish.

Table 11: Precision of household level EB model compared with direct estimates for super-areas

	Sri Lanka		Tanzania	
	Mean MSE	Mean CV	Mean MSE	Mean CV
Super-area estimates from direct sample estimate (Horvitz-Thompson approximation)	1.4	16.8	11.4	15.8
Area estimates from Household EB model	4.6	32.0	11.1	17.5

Notes: Top row shows the unweighted mean squared error (MSE) times 10,000 and coefficient of variation (CV) for non-monetary poverty estimates calculated from the national household budget survey across 25 Districts in Sri Lanka and 25 Regions in Tanzania. Bottom row shows the unweighted MSE times 10,000 and mean CV for non-monetary poverty estimates that incorporating geospatial data across 331 subdistricts in Sri Lanka and 159 Districts in Tanzania, benchmarked to survey estimates for 25 districts in Sri Lanka and 25 regions in Tanzania.

[30] In Sri Lanka, it is not possible to rescale the direct estimates to achieve 95 percent coverage because the area-level poverty and variance estimates are zero for more than 5 percent of the sub-districts.

V. Robustness checks for the household-level model

The previous section demonstrated that the household-level model produced more efficient estimates than the area level model and the direct estimates, while both the household level model and Fay-Herriot model both predicted poverty rates much more accurately than the direct estimates. The household-level model examined above was derived using a particular variant of the empirical best model, however, and it is informative to examine how the performance of the estimates varies based on different methodological choices.

Table 12 describes a set of eight robustness checks. The first two relate to the choice of software used to estimate the model. The baseline estimates were developed using a modified version of the R EMDI command that allows for weights in the simulation stage. In the first robustness check, this is compared with the latest version 3.0 of the Stata SAE package. This version allows for weights in both the model fitting and simulation stage, and implements the same bootstrap method as the R EMDI package. However, it uses a slightly different fitting method to estimate the model, adds a heteroscedasticity correction, and does not append the sample data to the census. The second robustness check uses version 2.0 of the Stata SAE package. This version of the Stata SAE package employs a traditional clustered bootstrap approach, which estimates the variance of the predicted poverty rates, instead of the parametric bootstrap approach used by the other packages to estimate mean squared error. The third robustness check does not use the empirical best method, to verify that the estimates are imprecise when only using synthetic predictions from the model. Version 2.0 of the Stata SAE package is used for this third robustness check because, unlike version 3 and the EMDI package, it allows the empirical best option to be disabled.

The fourth and fifth robustness checks experiment with disabling the heteroscedasticity correction in version 3 of the Stata package, and dropping the use of weights in the R EMDI package in the simulation stage. These are useful to check because, as noted above, the R EMDI package does not implement a heteroscedasticity correction, and the publicly available version does not allow for weights in the simulation stage. The sixth robustness check considers using a Box-Cox transformation rather than an Ordered Quantile Normalization to transform the dependent variable in the model. This gives an indication of how much the use of the Ordered Quantile Normalization improves the results and ensures that the estimates remain reasonable when using a more traditional normalization method. The seventh robustness check considers the predictions when using a stepwise procedure, set with a probability threshold of 0.01, is used to select the model instead of the plugin lasso method. The final robustness check is a simple test of the extent to which the results are robust to the prevailing poverty rate. This check swaps the country poverty rates, setting the Tanzanian poverty line at the 4^{th} percentile and the Sri Lankan poverty line at the 20^{th} percentile. These are then compared against the corresponding poverty rates calculated in the census using these new, swapped poverty rates.

Table 12: Description of robustness checks

Robustness check	Software package	Options
0. Baseline estimates	Modified R EMDI	Simulation stage weights, no heteroscedasticity correction. MLE fitting method, welfare transformed using ordered quantile normalization, point and MSE estimates benchmarked to survey super-areas, lasso model selection. Poverty rate of 4% in Sri Lanka and 20% in Tanzania.
1. Stata parametric bootstrap	Stata SAE version 3	Same as (0) but with sample weights, heteroscedasticity correction and H3 fitting method
2. Stata traditional bootstrap	Stata SAE version 2	Same as (1) with traditional non-parametric bootstrap instead of parametric bootstrap during simulation stage
3. No Empirical Best	Stata SAE version 2	Same as (2) with empirical best and bootstrap options disabled
4. Stata parametric bootstrap with no heteroscedasticity correction	Stata SAE version 3	Same as (1) with no heteroscedasticity correction
5. No weights	R EMDI	Same as (0) with no simulation stage weights
6. Box-Cox transformation	Modified R EMDI	Same as (0) but transform welfare using Box-Cox transformation with lambda selected to minimize skewness instead of ordered quantile normalization.
7. Stepwise model selection	Modified R EMDI	Same as (0) but select prediction model and heteroscedasticity model (for Stata SAE) using stepwise selection at p=0.01
8. Swapped poverty rates	Modified R EMDI	Same as (0) but with poverty rate of 20% in Sri Lanka and 4% in Tanzania

Table 13 displays the results of these eight robustness checks. For the purposes of brevity, we report the mean poverty, MSE, correlation, and coverage rate. [31] We begin by comparing the results produced by the baseline EMDI estimates with version 2 of the Stata SAE package, which implements the traditional bootstrap method, with and without the Empirical Best option. For all three estimates, the unweighted average of the predictions prior to benchmarking are close to the actual poverty rate in the census. The Stata package with the traditional bootstrap, however, gives significantly less precise estimates in both countries, as seen by the higher mean squared errors and CV. Purely synthetic predictions obtained without the empirical best option are substantially less precise than the empirical best predictions using the traditional clustered bootstrap in Sri Lanka, and moderately less precise in Tanzania. In both countries, however, the synthetic predictions are far less precise than estimates obtained using a parametric bootstrap approach, demonstrating the high efficiency cost of failing to incorporate sample

[31] Results for all areas are available upon request.

information when the auxiliary data are aggregated to the subarea level.[32] The estimates produced by the Stata SAE with the traditional bootstrap are less accurate than the baseline EMDI estimates in Sri Lanka, but more accurate in Tanzania. The coverage rate of the traditional bootstrap estimates is 93 percent, which is very close to the target 95 percent rate. This high coverage rate results from two countervailing biases in the variance estimate; the traditional bootstrap overstates the variance by failing to hold the composition of the sample constant across replications, which counteracts the downward bias due to the failure to account for the positive correlation among the sample residuals. Overall, the traditional bootstrap performs quite well in these two settings, in terms of accuracy and coverage, but poorly in terms of efficiency.

The estimates produced by the latest version of Stata SAE with the parametric bootstrap are slightly less accurate, in terms of its correlation with the full census, than those produced by the EMDI package.[33] The Stata SAE estimates are also less efficient than the baseline EMDI estimate in Sri Lanka, as the mean MSE rises from 4.6 to 6.8. In Tanzania, in contrast, the Stata SAE package with the parametric bootstrap yields estimates that are slightly more efficient than the baseline estimates, but because the estimates are less accurate, the coverage rate falls significantly from 74.8 to 64.2 percent.

The fifth row of the table presents results without the heteroscedasticity correction. Disabling the correction brings the mean poverty rate closer to the true 4 percent rate in Sri Lanka, but further from the assumed poverty rate of 20 percent in Tanzania. In Sri Lanka, disabling the correction increases the accuracy of the estimates to slightly above the baseline estimates. In Tanzania, disabling the heteroscedasticity correction has minor effects on accuracy and slightly decreases precision, improving the coverage rate. The limited effects of the heteroscedasticity correction likely results from the negligible explanatory power of the "alpha model" that predicts the variance of the residuals. The R^2 of the alpha model, which like the "beta" prediction model was selected using the plugin Lasso method to avoid overfitting, is only 0.004 in Sri Lanka and 0.003 in Tanzania.

The bottom four rows of the table present a variety of robustness checks using the EMDI package. The first of these uses the publicly available version of the package that does not allow for household weights when aggregating the simulated household poverty outcomes into an area poverty rate. The estimator without weights is slightly less accurate than those with weights in each country, with a corresponding mild decline in both efficiency and coverage. The second EMDI robustness check uses a more standard Box-Cox transformation of the welfare variable instead of the ordered quantile normalization transformation. This leads to considerable downward bias in the estimates in both countries, leading us to prefer the ordered quantile normalization, although the correlation and coverage rate changes little when using the Box-Cox transformation. The third EMDI robustness check reports the results of a model selected using stepwise regression, with a probability

[32] The penalty for failing to implement EB in a traditional bootstrap setting is smaller in Tanzania than Sri Lanka, although the estimates without EB are very imprecise in both countries. This is likely because the traditional bootstrap with EB performs gives less precise estimates in Tanzania than Sri Lanka, due to the smaller number of areas in Tanzania.

[33] This may result from the failure of the Stata SAE package to append the sample to the census when simulating poverty. The sample constitutes about 0.5 percent of Sri Lankan households while the Tanzanian sample only contains about 0.1 percent of households, which is consistent with the greater fall in accuracy in Sri Lanka.

Table 13: Robustness check results for alternative methods for household level model

	Sri Lanka					Tanzania				
	Mean poverty	Mean MSE	Mean CV	Corr	CR	Mean poverty	Mean MSE	Mean CV	Corr	CR
Baseline EMDI Estimates	3.7	4.6	32.0	87.6	84.3	18.6	11.1	17.5	87.6	74.8
Stata SAE with traditional bootstrap	3.7	10.9*	43.6	85.5	95.5	19.7	44.6*	33.4	88.3	93.1
With no Empirical Best	3.7	17.5*	62.4	83.6	99.1	19.8	55.9*	38.0	89.2	95.6
Stata SAE with parametric bootstrap	3.6	6.8	38.4	85.6	85.8	21.4	10.0	15.6	85.0	64.2
No heteroscedasticity correction	3.8	5.5	33.9	88.0	84.9	22.4	10.7	16.0	84.4	66.7
EMDI estimates with:										
No simulation weights	3.7	4.8	31.9	87.5	82.8	18.3	11.8	17.8	87.4	73.0
Box-Cox transformation	3.5	4.7	31.2	87.4	81.9	15.8	13.1	18.7	87.5	76.1
Stepwise (p=0.01) model selection	3.7	4.7	27.9	83.0	80.7	18.8	9.8	16.4	86.5	69.8
Swapped poverty rates	19.6	18.4	17.4	94.8	84.9	2.8	1.6	32.9	86.1	69.8

Notes: Rows represent different variants of household-level model as listed in Table 12. Columns represent mean poverty rates, average mean squared error times 10,000, mean coefficient of variation, correlation with actual census value (Corr), and coverage rate (CR). All means are unweighted across 331 subdistricts in Sri Lanka and 159 Districts in Tanzania. All results except for mean poverty reflect benchmarking to survey estimates for 25 districts in Sri Lanka and 25 regions in Tanzania. Asterisks indicate mean estimated variance instead of mean squared error.

value threshold of 1 percent. The resulting models are overfit. This leads to less accurate predictions than the Lasso-selected model in both countries and slightly reduces mean squared error in Tanzania, leading to a moderate fall in coverage rates in both countries. Finally, we report the results of the EMDI estimators when the poverty rates of the two countries are swapped. Not surprisingly, the performance of the model improves in Sri Lanka when the poverty rate is set to 20 percent, as the remote sensing indicators better predict poverty rates when they are larger. When setting the poverty rate to 4 percent in Tanzania, the EMDI estimator tends to underestimate poverty, with a mean predicted poverty rate of 2.8 percent. However, the correlation between the estimates and the true census poverty rate remains strong at 86.1 percent, and the benchmarked coverage rate of 70 percent is respectable.

Overall, of the many robustness checks considered so far, there are only two main cases where the results are particularly sensitive to the choice of method. The first is the use of the traditional rather than parametric bootstrap, which leads to a much higher estimated variance, coefficient of variation, and coverage rate of the estimates. The traditional bootstrap is not recommended when using Empirical Best models on theoretical grounds, since the estimate is conditioned on the sample (Diallo and Rao, 2018). The other case where the results change dramatically is the omission of the empirical best option, which roughly quadruples the MSEs and doubles the CVs compared to the baseline estimates,

and yields substantially less accurate estimates in Sri Lanka as measured by the correlation with the census. Otherwise, the results are largely robust to the range of robustness checks considered.

Next, we perform two additional robustness checks. The first of these examines the results of the household model using different benchmarking procedures. In particular, we examine two alternatives to the baseline estimates, in which both the point estimates and standard errors are benchmarked. The first alternative only benchmarks the point estimates, while the second performs no benchmarking. The results are shown in Table 14. In both countries, not benchmarking the standard errors slightly reduces the average mean squared error. Average CV is minimized in both countries by benchmarking only the point estimates. In Sri Lanka, benchmarking only the point estimates slightly reduces the CV while not benchmarking increases it because an additional district is included.[34] This leads to a modest fall in coverage in Sri Lanka, but a more significant reduction in Tanzania, where the coverage rate declines from 74.8 percent with benchmarking to 69.8 percent without. Benchmarking the mean squared errors, all else equal, overstates uncertainty in cases like these where mean poverty is underestimated. This partially mitigates the downward bias in the household model's mean squared error estimates, which leads the benchmarking procedure to improve coverage. The benchmarked estimates are therefore preferred, although the results are qualitatively similar when either the point estimates are benchmarked, or when no benchmarking is carried out.

The final robustness check examines how the baseline household level model fares when random noise is added to the welfare aggregate. This better approximates a monetary welfare measure such as per capita consumption or income, which contain a greater portion of unexplained variance resulting from both transient welfare shocks and measurement error. We add two error term components to the existing measure of non-monetary welfare, to make an alternative noisier measure of "true welfare". The first error component is an area effect, drawn form a normal distribution with mean zero and variance 0.5, and the second is a household specific error term with mean zero and variance 1. These variance values were chosen arbitrarily, with the aim of achieving a model R^2 of approximately 0.2, which is similar to the R^2 when regressing monetary welfare on similar geospatial indicators in Tanzania (Belghith et al, 2020). The poverty line is set according to the new, noisier, welfare measure, equal to the 4th percentile of the sample welfare in Sri Lanka, and the 20th percentile of the sample in Tanzania.

Table 15 displays the results when estimating area-level poverty using direct estimates, the Fay-Herriot model, and the household model with a noisier welfare aggregate. Household model diagnostics are displayed below the results. In the household model, the additional noise causes the marginal R^2 to fall to about 0.2 in both cases. This is substantially lower than the values reported when predicting the non-monetary welfare index, which were 0.27 for Sri Lanka and 0.3 for Tanzania (Table 7). However, the conditional R^2 remains high, and even slightly above what is reported for the non-monetary welfare measure in Table 7. This reflects the additional variance in the area-specific component, which is captured by the mean of the sample residuals for each area and therefore incorporated into the small area predictions by the empirical best estimator.

[34] The estimated poverty rate for Ratnapura district is zero in the HIES sample. The CV is therefore not defined for this district after benchmarking and excluded from the average.

Table 14: Robustness check for benchmarking

	Sri Lanka					Tanzania				
Benchmarking	Mean poverty	Mean MSE	Mean CV	Corr	CR	Mean poverty	Mean MSE	Mean CV	Corr	CR
Baseline estimates	3.7	4.6	32.0	87.6	84.3	18.6	11.1	17.5	87.6	74.8
Benchmark point estimates only	3.7	4.2	29.9	87.6	79.5	18.6	9.7	16.9	87.6	71.1
No benchmarking	3.7	4.2	39.1	88.1	82.2	18.6	9.7	17.5	83.7	69.8

Notes: Columns represent mean poverty rates, mean squared error times 10,000, coefficient of variation, correlation with actual census value (Corr), and coverage rate (CR). Top row reports estimates when point estimates and MSEs are benchmarked to survey results for each super-area. Bottom row shows results without benchmarking.

The overall results, with a few exceptions, are similar to the main results for the non-monetary welfare measure reported in Tables 8 and 9. The household level model produces the most precise estimates, as judged by mean squared error. The MSE reduction achieved using the household-level model in Sri Lanka is smaller than the baseline results, reflecting the challenge of accurately predicting the bottom tail of the welfare distribution with a noisier welfare measure. The average CV rises greatly in Sri Lanka, due to large positive outliers, when using the noisier measure. These outliers are areas with very low predicted poverty rates, yet high mean squared error, due to the greater variability in the area effect in these simulations. This illustrates that average CV should be interpreted with caution in settings with low poverty rates and highly variable area effects. Nonetheless, even with a noisier welfare aggregate, the reduction in mean squared error Sri Lanka compared with the direct estimates is roughly equivalent to doubling the size of the sample. In Tanzania, where the poverty rate is much higher, the household-level model leads to a 78 percent reduction in the mean squared error, a factor comparable to that reported in Table 8. The analogous reduction in average CV is 30 percent, which is also comparable to roughly doubling the sample size.

Turning to accuracy and coverage, exploiting sub-area variation using the household model substantially improves accuracy, as measured by the correlation with the true noisier welfare measure. In Sri Lanka, the correlation is over 4 percentage points higher for the household level model than the Fay-Herriot model, while in Tanzania the comparable difference is over 10 percentage points. The correlation in absolute terms remains respectable in Sri Lanka (0.80) and high in Tanzania (0.84 percent) when using the noisier welfare measure. Finally, coverage rates for the household level model decline only modestly in Tanzania and increase significantly in Sri Lanka in comparison with the direct estimates. In Sri Lanka, the coverage rate reported in Table 15 for the noisy welfare measure is remarkably similar to the one for non-monetary poverty reported in Table 8 (84.9% vs. 83.7%). In Tanzania, the coverage rate is higher for the noisy welfare measure (78.0% vs. 73.6%). Accuracy and coverage are particularly important measures to consider because, unlike estimated precision, they are only observed when a "true" census population is available. Overall, the results presented in Table 15 indicate that the accuracy and coverage of the household level model remain high, in comparison with both the Fay-Herriot and direct estimates, when additional noise is introduced into the welfare measure.

Table 15: Robustness check for noisy welfare measure

	Sri Lanka					Tanzania				
	Mean poverty	Mean MSE	Mean CV	Corr	CR	Mean poverty	Mean MSE	Mean CV	Corr	CR
Direct estimates (H-T approximation)	4.0	15.9	65.3	59.0	74.6	20.0	82.3	45.4	77.2	83.0
Area level Fay-Herriot model	4.0	10.5	54.1	75.9	80.7	16.8	53.2	39.5	72.8	88.1
Household-level EB model	3.7	7.3	68.1	79.5	84.9	17.6	18.1	31.7	85.2	79.9
Household Model Diagnostics										
Marginal R^2	0.196					0.208				
Conditional R^2	0.304					0.329				
Skewness of area effect	0.14					-0.04				
Kurtosis of area effect	2.99					2.61				
Wilks-Shapiro P-value	0.40					0.33				
Skewness of household error	-0.03					-0.06				
Kurtosis of household error	3.12					3.08				
Variance of estimated area effect	0.110					0.122				
Variance of estimated household residual	0.702					0.671				
Percentage of variance of error term due to area effect	15.7%					18.2%				

VI. Simulations

The results presented in the preceding two sections are based on small area estimation using one sample of households for each country.[35] The samples that were used are particularly significant, since they reflect the actual enumeration areas sampled by the National Statistics Offices for their household

[35] MSE in the baseline estimates were estimated by simulating different populations while maintaining the same sample composition in each parametric bootstrap replication.

budget surveys. Nonetheless, it is useful to check that the strong performance of the household-level model holds when estimated in simulation settings, as is commonly done in the literature.[36]

a. Model-based simulation

We begin by conducting a Monte Carlo simulation study, following the basic structure outlined in Das and Chambers (2017). We perform 100 simulations, drawing a new simulated population each time, in 80 areas. Each of these areas contains between 15 and 29 subareas, with the exact number chosen randomly along a uniform distribution. Each subarea contains 15 and 29 households, also chosen randomly from a uniform distribution. Each household is assumed to contain one member.

At the sub-area level, we draw two X variables, $X1_{sa}$ and $X2_{sa}$, from a joint normal distribution. All households in the same sub-area are assigned the same random values of $X1_{sa}$ and $X2_{sa}$. $X1_{sa}$ is distributed with mean 0 and variance 2. The mean of $X2_{sa}$ is set equal to 0.05 times the area index, which varies from 1 to 80, with a variance of 2. The covariance of $X1_{sa}$ and $X2_{sa}$ is set to 0.5. We draw area effects (η_a), subarea effects (u_{sa}), and an idiosyncratic error term for each household (ε_i). These error terms are drawn from normal distributions with mean zero and variances of 1 for the area effect, 0.5 for the sub-area effect, and 2 for the household idiosyncratic error. We then generate the population welfare measure as

$$(6) \qquad y_i = 6 + 0.5 * X1_{sa} - 0.6 * X2_{sa} + \eta_a + u_{sa} + \varepsilon_i$$

The variance structure of the two X variables and variance components was selected to achieve an approximate R^2 of 0.25 when regressing household welfare on $X1_{sa}$ and $X2_{sa}$.

From this population, we set the poverty line at the 25th percentile of the welfare distribution and calculate population poverty rates for each area. We then draw a two-stage sample of 3 subareas per area in the first stage and 10 households per sampled subarea in the second. We construct sample weights as the inverse of the probability of being sampled, and take the weighted 25th percentile of the sample distribution of welfare as the poverty line. Finally, we estimate a small area estimation model in the sample using the modified EMDI package, utilizing the population values of $X1_{sa}$ and $X2_{sa}$ to simulate welfare for all households in the population. The resulting simulated welfare aggregate is then compared with the poverty line derived from the sample, and averaged over households for each area to obtain area headcount poverty rates and estimated mean squared errors.[37] This entire process, starting with generating the simulated population, is then repeated 100 times. Therefore, the final output from the procedure is a set of 8,000 population poverty rates and small area estimates, derived from 100 simulations of 80 areas.

Table 16 displays the results. The small area estimates reduce the mean squared error by a factor of 10 and average CV by a factor of roughly 3 compared with the traditional EA-clustered direct estimates, while maintaining a high coverage rate close to 95 percent. The mean poverty is nearly exactly 25 percent, and the correlation compared with poverty rates in the population increases from 0.84 for the direct estimates to 0.98 for the small area estimates. The root mean squared error of the predictions

[36] Evaluations of small area estimation methods using model and/or design-based simulations include Das and Chambers (2017), Demombynes et al (2007), Elbers, Lanjouw and Leite (2008), Tarozzi and Deaton (2009), Tarozzi (2011) and Tzavidis et al (2018), among many others.

[37] The results are not benchmarked to super-areas, because super-areas are not defined in this simulation.

falls by a factor of three. The household level model performs well in this stylized setting because the residuals are distributed normally by construction, and a large percentage of variation in simulated welfare is explained by the two predictor variables. It is reassuring, though, that the small area estimator performs well in a favorable controlled environment with populations that vary across simulations.

b. Design-based simulation

While model-based simulations are useful to explore the performance of the estimator in different populations, the results of model-based simulations inevitably depend on the set of assumptions used to construct the simulated data. To verify that the household-level model works well when estimated in multiple samples, this section describes the results of a design-based simulation. To reduce computation time, the simulation is limited to three regions in Northeast Tanzania near the Kenyan border: Arusha, Kilimanjaro, and Tanga. Between them, these three regions according to the census contain 24 districts, 2,078 villages, 10,964 enumeration areas, and nearly 1.17 million households.[38] The 2018 Household Budget Survey sampled 1,217 households in 97 enumeration areas from these three regions, for an average sample of 12.5 households per enumeration area. The three regions are disproportionately poor, as the non-monetary poverty rate is 46.5 percent when using a poverty line set at the 20th percentile of the national distribution.

To conduct the design-based simulations, we repeat the following seven-step procedure 250 times:

1. Draw a simple random sample of 4 EAs for each of the 24 districts from the census, which selects a total of 96 EAs from the population of 10,964 census EAs contained in the three regions.

2. Draw a simple random sample of 13 households for each selected EA to generate a sample of 1,248 households.

3. Construct sample weights for each household equal to the inverse of their selection probability.[39]

4. Use the ordered quantile normalization to construct normalized non-monetary welfare in the sample.

5. Using the household sample from step 2, the sample weights from step 3, and the normalized welfare measure from step 4, estimate a lasso regression of normalized non-monetary welfare on all candidate geospatial indicators.[40]

6. Using the same inputs as the previous step, calculate the 46.5th percentile of the sample distribution of normalized non-monetary welfare to use as the poverty line.

[38] We selected three regions to use for this exercise because the lasso procedure did not select any variables when using only one or two regions, when mimicking the sample size of the budget survey. We selected these particular three regions because they are disproportionately poor and because they were the first three regions in numeric order following the capital of Dodoma.

[39] The household sample weights are the product of two ratios: The number of census EAs in each district divided by 4, and the number of households in each sampled EA divided by 13. These are multiplied by household size to obtain population weights.

[40] As before, we use the plugin method for selecting lambda in the lasso.

Table 16: Results of model-based simulations

	Monte Carlo simulations						
	Mean Poverty	Mean MSE	Mean CV	Corr	ARB	RMSE	Coverage rate
Direct survey estimates							
H-T approximation	24.9	51.8	31.5	0.840	-0.005	0.060	95.7
EA-Clustering	24.9	44.4	29.8	0.840	-0.005	0.060	94.0
Household-level EB model	25.1	4.7	9.2	0.978	0.012	0.020	96.4
Model diagnostics							
Marginal R^2			0.236				
Conditional R^2			0.237				
Skewness of area effect			0.080				
Kurtosis of area effect			3.00				
Wilks-Shapiro P-value of area effect			0.481				
Skewness of household error			-0.002				
Kurtosis of household error			3.00				
Wilks-Shapiro P-value of household error			0.483				

Notes: Results reflect averages across 8000 estimates, representing 100 simulations of 80 simulated areas. Diagnostic results reflect averages over 100 simulations. H-T indicates Horvitz-Thompson approximation of variance.

7. Use the modified EMDI package to generate point and MSE estimates for each district based on the census data of 1.17 million households containing all model prediction variables and household size for these regions. The model estimates derived from the sample are not weighted, and the dependent variable is normalized welfare calculated in step 4. Census households are weighted by household size when estimating district level headcount poverty rates. As above, we use 100 Monte Carlo simulations to generate point estimates of headcount poverty and 100 bootstrap replications to estimate the mean squared error.

After repeating this seven-step process 250 times, we can calculate average evaluation statistics of the type reported in Tables 7 and 8 across the 24 districts and 250 simulations.[41] Since the Fay-Herriot

[41] Poverty rates for each district are calculated based on a poverty line set at the 46.5th percentile of the national distribution of non-monetary welfare.

model would not be expected to perform well in a sample of only 24 districts, we only report average evaluation statistics for the two types of direct estimates and the household level unit level models.

Results for the household-level model are reported with and without benchmarking. As above, the benchmarked point estimates were rescaled to match the direct estimates for each of the three regions in the population, for each sample drawn in each simulation, and the mean squared error were multiplied by the square of the same scaling factor. The direct estimates are computed in each simulation by calculating the weighted share of sample households that are poor in each district, using the weights calculated in step 3 above. We again compare the direct estimates obtained using the Horvitz-Thompson approximation with results from standard enumeration area clustering, which give identical point estimates but differ in their estimated variances.

Table 17 displays the results, which confirm that the household model continues to perform well when using repeated samples from the three selected regions. The household model in these simulations demonstrate a moderate amount of bias, with a mean of 42.4 percent prior to benchmarking as compared with 46.5 percent in the population. This bias results from the distribution of the residuals deviating substantially from normality at the tails.[42]

The estimates produced by the household level model, however, are far more precise than the Horvitz-Thompson direct estimates, with average mean squared errors that are about a quarter as large and average CVs about half the size. The household model estimates are only moderately more precise than the standard direct estimates, reflecting a high degree of correlation in household welfare across EAs within district. The standard direct estimates ignore this correlation and therefore greatly underestimate uncertainty, which are reflected in a coverage rate of 48.7 percent. In contrast, the coverage rate of the benchmarked estimates from the household model is 68 percent, which is moderately below the coverage rate for all of Tanzania using the HBS sample reported in Table 8 (74.8%). The coverage rate of the household model, despite being much more precisely estimated, is only modestly below the Horvitz-Thompson direct estimates. This is because the household model estimates are more accurate, as seen in the higher correlation and lower root mean squared error.

To sum up, when using repeated samples in three poor Tanzanian regions, incorporating remote sensing data in a household-level model dramatically improves precision compared to the Horvitz-Thompson direct estimates. Adding the remote sensing data reduces mean squared error by about 75 percent and CV by about half, which is more or less comparable to quadrupling the sample size. Because the small area estimates are much more accurate than the direct estimates, the increased precision comes at a modest cost in terms of coverage rate. Compared with the standard direct estimates assuming independence across enumeration areas, the small area estimates are not only slightly more precise, but also significantly boost accuracy and coverage. Overall, the main results derived for the full country based on budget survey the specific sample drawn for the budget survey are maintained when focusing on three Tanzanian regions and generating estimates of the area poverty rates and their mean squared errors using 250 different samples.

[42] As shown in the bottom of Table 17, the kurtosis of the residuals is 2.5 for the area effect and 3.7 for the household errors.

Table 17: Results of Design-Based Simulation

	Arusha, Kilimanjaro, and Tanga Regions						
	Mean Poverty	ARB	Mean MSE	Mean CV	Corr	RMSE	Coverage rate
Direct survey estimates							
H-T approximation	47.0	-0.031	135.4	29.9	0.673	0.212	72.1
EA-Clustered standard errors	47.0	-0.031	52.0	21.0	0.673	0.212	48.7
Household-level EB model							
Benchmarked	45.7	0.028	29.9	14.0	0.828	0.112	68.0
Non-benchmarked	42.4	-0.046	26.0	14.0	0.803	0.083	66.6
Model diagnostics							
Marginal R^2	0.324						
Conditional R^2	0.361						
Skewness of area effect	0.13						
Kurtosis of area effect	2.45						
Wilks-Shapiro P-value of area effect	0.59						
Skewness of hh error	-0.19						
Kurtosis of hh error	3.66						
Wilks-Shapiro P-value of hh error	0.002						

Notes: Results reflect averages across 6000 estimates, representing 250 simulations of 24 districts. Fay-Herriot model not included due to small number of districts in sample. Diagnostic results reflect averages over 250 simulations

VII. Conclusion

Policy makers worldwide are increasingly demanding more granular data to inform decisions. This paper examines the extent to which supplementing a synthetic sample drawn from the census with geospatial data at the subarea level in Sri Lanka and Tanzania generates more precise and accurate estimates of non-monetary poverty. The paper employs empirical best predictor models because they are more accessible to most statisticians in National Statistics Offices than hierarchal Bayesian models. In addition, empirical best models have a long history in small area estimation, are similar to other commonly used methods, and have been implemented in multiple publicly available software packages. The analysis examines non-monetary poverty in order to evaluate the estimates against census data, which sheds light on the performance of different methodological approaches and software packages.

The results are encouraging. The correlation between the small area estimates and the actual census is roughly 88 percent in both countries, and substantially greater than the analogous correlations for the direct survey estimates, which are 73 and 77 percent in Sri Lanka and Tanzania, respectively. The household-level model, compared with standard clustered survey estimates, reduces the mean squared error by about two-thirds in Sri Lanka and four-fifths in Tanzania, which is roughly equivalent to tripling

and quintupling the effective sample size. The results are robust to alternative implementation options, adding additional noise to the welfare measure, and the implementation of both a model-based Monte Carlo simulation, and a design-based simulation from three regions in Northeast Tanzania. When rescaling mean squared errors to equalize coverage rates, the household-level model remained more efficient than a Fay-Herriot model, especially in Sri Lanka. The strong performance of the household-level EBP model, in comparison with the area-level Fay-Herriot model, is consistent with simulation results reported in Hidiroglou and You (2016). That study concluded, on the basis of a model-based simulation study with a linear model and unit-level auxiliary data, that unit-level models achieve lower bias and higher coverage rates than area-level models.

The household-level EBP model moderately underestimates uncertainty, leading to sizeable drops in coverage rates compared with estimates from the area-level Fay-Herriot model, especially in Tanzania. This occurs because the model does not incorporate sample weights, and assumes that survey data are independent within area when conditioning the random effect on the sample residuals. In addition, both the household and area-level models do not account for uncertainty in the estimated variance components in the model. Nonetheless, the coverage rate of the household model in both cases is comparable to standard survey estimates that cluster on enumeration areas. Furthermore, the household model performs significantly better than the Fay-Herriot model in terms of accuracy and efficiency when additional noise is added to the welfare measure, especially in Tanzania.

The financial cost of this type of small area estimation procedure is generally low and falling rapidly. Much of the auxiliary data used for the small area estimation, such as estimates of built-up area, nighttime lights, and vegetation, are freely available. There are two notable exceptions. First, the calculation of spatial features at a national scale requires constructing a cloud-free mosaic of Sentinel imagery, considerable computing power, and the expertise to implement software to calculate spatial features. Second, the data on Tanzanian building footprints are proprietary. However, data on building footprints are increasingly being released in the public domain and it is not difficult to envision information on building footprints becoming freely available in the coming years, potentially through Open Street Map as it continues to improve in accuracy and coverage. Finally, access to subarea survey identifiers and shapefiles, or access to EA geocoordinates, is necessary to link survey data to geospatial indicators. This is not feasible in all contexts, but the growing popularity of geospatial analysis and CAPI data collection is making geospatial survey analysis more common. A conservative estimate is that the time and expertise required to generate these types of estimates costs $50,000 to $150,000. This is minor compared to the value created by even doubling the effective sample size of nationally representative household surveys that often cost at least a million dollars to field.

The results could be improved by further refinements in software and methods. One avenue for further research is to explore the performance of a subarea-level model, an extension of the Fay-Herriot model specified at the subarea level with an area-level mixed effect (Torabi and Rao, 2015). Estimating such a model at the subarea level makes it easier to properly account for sample design effects. However, modeling poverty rates as a linear function of predictors may generate less accurate estimates, especially when poverty rates are low. It would be useful to better understand how the results of a properly specified subarea model compare with a household-level model using subarea predictors.

A second line of research could focus on improving the household-level model by estimating uncertainty more accurately. The estimated variance of the area random effect, for example, could be adjusted to

account for sample design effects among the residuals, which would increase the estimated mean squared errors and improve coverage rates.[43] In addition, benchmarking could be incorporated directly into the bootstrap procedure, to obtain more accurate estimates of benchmarked mean squared errors. Third, it should be fairly straightforward to incorporate the ordered quantile normalization transformation directly to the R EMDI package, which already allows for Box-Cox and log shift transformations. This would enable the software to estimate poverty gaps and severity in addition to headcount rates when using the ordered quantile normalization transformation.

This study only considered a measure of non-monetary poverty, in order to evaluate the results of the area predictions against actual census data. This raises the important question of the extent to which the results generalize to predicting monetary poverty rates. The types of geospatial indicators considered in this study are moderately less predictive of monetary welfare than non-monetary welfare. For example, when predicting district-level monetary poverty in Tanzania using the same geospatial indicators considered above, the geospatial variables selected by a stepwise procedure explained 23.5 percent of the variation in household per capita consumption, versus 30 percent for a lasso-selected model of non-monetary welfare.[44] This reflects the difficulty of predicting both transient household welfare shocks and measurement error, both of which are present to a greater degree in measures of monetary poverty than non-monetary poverty. It is encouraging that the household-level EBP model continues to perform well when noise is added to non-monetary welfare. The efficiency gain from using the household model was notably smaller in Sri Lanka when using the noisier welfare measure, but even in that context incorporating geospatial data led to an increase in precision roughly equivalent to doubling the sample size. These results are only suggestive, however, and an important outstanding research agenda is to compare small area estimates of monetary poverty obtained using geospatial data to those obtained from standard census-based poverty maps directly from census data. This can confirm that the encouraging results presented above apply to monetary poverty.

A final caveat, which applies to small area estimation in general, is the challenge of measuring local transient shocks. In the context of geospatial data, key indicators used to generate model predictions, such as building counts, are not necessarily affected by local economic shocks. This also poses an issue for traditional census-based poverty maps, as the census predictions tend to rely heavily on indicators such as educational attainment and household size that may also fail to track local shocks. Empirical Best methods partly address this issue by incorporating welfare information from the sample. Nonetheless, future research is needed to explore which geospatial indicators better reflect local economic shocks.

Overall, the results from an evaluation exercise using data from two countries demonstrate major efficiency gains from combining survey data with geographically comprehensive geospatial data to generate small area estimates of non-monetary poverty. These efficiency gains came at no cost to coverage in Sri Lanka and a moderate cost in coverage rates in Tanzania, and the latter can be addressed through further refinements to the methodology. The financial cost of incorporating geospatial data is low, relative to the cost of achieving similar efficiency gains by surveying additional households. While there is room for further methodological improvement, these techniques can currently be applied at modest cost to generate more granular and informative estimates of non-monetary welfare.

[43] Chambers and Das (2017) propose a similar strategy of adjusting traditional ELL estimates to correct for biased estimates of mean squared error.

[44] See Belghith et al, 2020. This corresponds to the marginal R2 reported in Table 6.

References

Ayush, K., Uzkent, B., Burke, M., Lobell, D., & Ermon, S. (2020). Generating Interpretable Poverty Maps using Object Detection in Satellite Images. arXiv preprint arXiv:2002.01612.

Bartlett, Maurice S. "The use of transformations." *Biometrics* 3.1 (1947): 39-52.

Battese, George E., Rachel M. Harter, and Wayne A. Fuller. "An error-components model for prediction of county crop areas using survey and satellite data." Journal of the American Statistical Association 83.401 (1988): 28-36.

Belghith, Nadia Belhaj Hassine; Karamba, R. Wendy; Talbert, Elizabeth Ann; De Boisseson, Pierre Marie Antoine. 2020. Tanzania - Mainland Poverty Assessment 2019 : Part 1 : Path to Poverty Reduction and Pro-Poor Growth (English). Washington, D.C. : World Bank Group.

Belloni, Alexandre, and Victor Chernozhukov. "High dimensional sparse econometric models: An introduction." *Inverse Problems and High-Dimensional Estimation*. Springer, Berlin, Heidelberg, 2011. 121-156.

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

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies* 81.2 (2014): 608-650.

Boonstra, H. J. "hbsae: Hierarchical Bayesian small area estimation." *R package version* 1 (2012).

Box, George EP, and David R. Cox. "An analysis of transformations." *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2 (1964): 211-243.

Chandra, H., Sud, UC. and Gupta V.K. (2013). Small Area Estimation under Area Level Model Using R Software

Corral Rodas, Paul, William Seitz, João Pedro Azevedo, and Minh Cong Nguyen. "FHSAE: Stata module to fit an area level Fay-Herriot model." (2018).

Corral Rodas, Paul Andres, Isabel Molina, and Minh Cong Nguyen. "Pull Your Small Area Estimates up by the Bootstraps." *World Bank Policy Research Working Paper* 9256 (2020).

Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE, 2005.

Das, Sumonkanti, and Ray Chambers. "Robust mean-squared error estimation for poverty estimates based on the method of Elbers, Lanjouw and Lanjouw." Journal of the Royal Statistical Society: Series A (Statistics in Society) 180.4 (2017): 1137-1161.

Demombynes, G., Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2007). How good a map? Putting small area estimation to the test. The World Bank.

Diallo, Mamadou S., and J. N. K. Rao. "Small area estimation of complex parameters under unit-level models with skew-normal errors." Scandinavian Journal of Statistics 45.4 (2018): 1092-1116.

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

Elbers, Chris, Peter Lanjouw, and Phillippe George Leite. Brazil within Brazil: Testing the poverty map methodology in Minas Gerais. The World Bank, 2008.

Engstrom, Ryan, Jonathan Hersh, and David Newhouse. "Poverty from space: Using high-resolution satellite imagery for estimating economic well-being." (2017).

Engstrom, Ryan, David Newhouse, and Vidhya Soundararajan (2019a). Estimating Small Area Population Density Using Survey Data and Satellite Imagery: An Application to Sri Lanka. The World Bank, 2019.

Engstrom, Ryan, Pavelsku, Dan, Tomomi, Tanaka, Wambile, A. (2019b). Mapping Poverty and Slums Using Multiple Methodologies in Accra , Ghana. Joint Urban Remote Sensing Conference 2019, 1–4.

Engstrom, R, Harrison, R., Mann, M., & Fletcher, A. (2019c). Evaluating the relationship between contextual features derived from very high spatial resolution imagery and urban attributes: A case study in Sri Lanka. 2019 Joint Urban Remote Sensing Event, JURSE 2019. https://doi.org/10.1109/JURSE.2019.8809041

Erciulescu, Andreea L., Nathan B. Cruze, and Balgobin Nandram. "Model-based county level crop estimates incorporating auxiliary sources of information." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182.1 (2019): 283-303.

Fay III, Robert E., and Roger A. Herriot. "Estimates of income for small places: an application of James-Stein procedures to census data." Journal of the American Statistical Association 74.366a (1979): 269-277.

González-Manteiga, Wenceslao, et al. "Bootstrap mean squared error of a small-area EBLUP." *Journal of Statistical Computation and Simulation* 78.5 (2008): 443-462.

Guadarrama, María, Isabel Molina, and J. N. K. Rao. "A comparison of small area estimation methods for poverty mapping." *Statistics in Transition new series* 1.17 (2016): 41-66.

Haslett, Stephen J. "Small area estimation using both survey and census unit record data." *Analysis of Poverty Data by Small Area Estimation* (2016): 325-348..

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Head, Andrew, et al. "Can human development be measured with satellite imagery?." *ICTD*. 2017.

Hidiroglou, Michael A., and Yong You. "Comparison of unit level and area level small area estimators." *Survey Methodology* 42.41-61 (2016).

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663-685

Huang, Xin, Liangpei Zhang, and Pingxiang Li. "Classification and extraction of spatial features in urban areas using high-resolution multispectral imagery." *IEEE Geoscience and Remote Sensing Letters* 4.2 (2007): 260-264.

James, W., and C. Stein. "Proc. Fourth Berkeley Symp. Math. Statist. Probab." *Estimation with quadratic loss*. Vol. 1. Univ. California Press, 1961. 361-379.

Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." *Science* 353.6301 (2016): 790-794.

Kreutzmann, Ann-Kristin, et al. "The R package emdi for the estimation and mapping of regional disaggregated indicators." *Journal of Statistical Software* (2018).

Hall, Peter, and Tapabrata Maiti. "On parametric bootstrap methods for small area prediction." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.2 (2006): 221-238.

Jarvis, A., et al. "SRTM 90m Digital Elevation Database v4. 1." (2008).

Jiang, Jiming, and Partha Lahiri. "Mixed model prediction and small area estimation." *Test* 15.1 (2006):

Halbmeier, Christoph, et al. "The fayherriot command for estimating small-area indicators." *The Stata Journal* 19.3 (2019): 626-644.

Kottek, Markus, et al. "World map of the Köppen-Geiger climate classification updated." *Meteorologische Zeitschrift* 15.3 (2006): 259-263.

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

Marhuenda, Yolanda, et al. "Poverty mapping in small areas under a twofold nested error regression model." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.4 (2017): 1111-1136.

Matsuura, Kenji, and Cort J. Willmott. "Terrestrial precipitation: 1900–2017 gridded monthly time series." *Electronic. Department of Geography, University of Delaware, Newark, DE* 19716 (2018).

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

Miller, Rupert G. "The jackknife-a review." *Biometrika* 61.1 (1974): 1-15.

Molina, Isabel, and J. N. K. Rao. "Small area estimation of poverty indicators." *Canadian Journal of Statistics* 38.3 (2010): 369-385.

Molina, Isabel, and Yolanda Marhuenda. "sae: An R package for small area estimation." *The R Journal* 7.1 (2015): 81-98.

Morris, Carl N. "Parametric empirical Bayes inference: theory and applications." *Journal of the American Statistical Association* 78.381 (1983): 47-55.

Myint, Soe W., Victor Mesev, and Nina Lam. "Urban textural analysis from remote sensor data: Lacunarity measurements based on the differential box counting method." *Geographical Analysis* 38.4 (2006): 371-390.

Nguyen, Minh C., et al. *Small Area Estimation: An extended ELL approach*. mimeo, 2017

Pesaresi, Martino, Andrea Gerhardinger, and François Kayitakire. "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 (2008): 180-192.

Peterson, Ryan A., and Joseph E. Cavanaugh. "Ordered quantile normalization: a semiparametric transformation built for the cross-validation era." *Journal of Applied Statistics* (2019): 1-16.

Pfeffermann, Danny, Anna Sikov, and Richard Tiller. "Single-and two-stage cross-sectional and time series benchmarking procedures for small area estimation." *Test* 23.4 (2014): 631-666.

Pokhriyal, Neeti, and Damien Christophe Jacques. "Combining disparate data sources for improved poverty prediction and mapping." *Proceedings of the National Academy of Sciences* 114.46 (2017): E9783-E9792.

Prasad, NG Narasimha, and Jon NK Rao. "The estimation of the mean squared error of small-area estimators." *Journal of the American statistical association* 85.409 (1990): 163-171.

Sandborn, Avery, and Ryan N. Engstrom. "Determining the relationship between census data and spatial features derived from high-resolution imagery in Accra, Ghana." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.5 (2016): 1970-1977.

StataCorp, L. P. "Stata LASSO reference manual." *College Station, TX: StataCorp LP* (2019).

Steele, Jessica E., et al. "Mapping poverty using mobile phone and satellite data." *Journal of The Royal Society Interface* 14.127 (2017): 20160690.

Tak, Hyungsuk, Joseph Kelly, and Carl N. Morris. "Rgbp: an R package for gaussian, poisson, and binomial random effects models with frequency coverage evaluations." *arXiv preprint arXiv:1612.01595* (2016).

Tarozzi, Alessandro, and Angus Deaton. "Using census and survey data to estimate poverty and inequality for small areas." *The review of economics and statistics* 91.4 (2009): 773-792.

Tarozzi, Alessandro. "Can census data alone signal heterogeneity in the estimation of poverty maps?." *Journal of Development Economics* 95.2 (2011): 170-185.

Torabi, Mahmoud, and J. N. K. Rao. "On small area estimation under a sub-area level model." *Journal of Multivariate Analysis* 127 (2014): 36-55.

Tzavidis, Nikos, et al. "From start to finish: a framework for the production of small area official statistics." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4 (2018): 927-979.

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.

Unsalan, C., and Kim L. Boyer. "Classifying land development in high-resolution panchromatic satellite images using straight-line statistics." *IEEE Transactions on Geoscience and Remote Sensing* 42.4 (2004): 907-919.

Van der Waerden, B. L. "Order tests for the two-sample problem and their power." *Indagationes Mathematicae (Proceedings)*. Vol. 55. North-Holland, 1952.

Wang, Junyuan, Wayne A. Fuller, and Yongming Qu. "Small area estimation under a restriction." *Survey methodology* 34.1 (2008): 29.

Wood-Sichra, Ulrike, Alison B. Joglekar, and Liangzhi You. *Spatial Produc‐tion Allocation Model (SPAM) 2005: Technical Documentation*. HarvestChoice Working Paper, 2016.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., ... & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications*, *11*(1), 1-11.

Zhao, Qinghua. "User manual for povmap." *World Bank. http://siteresources. worldbank. org/INTPGI/Resources/342674-1092157888460/Zhao_ ManualPovMap. pdf* (2006).

Annex A: Software packages for household-level EBP models

As noted in the text, there are five main differences between the different packages, which are listed in Table 4. The first important difference is that the stata SAE package implements a heteroscedasticity correction, following Elbers, Lanjouw and Lanjouw (2003). This correction estimates an "alpha model" that regresses the squared residuals from an OLS regression on subarea characteristics to obtain an estimate of the variance. The estimated variance is then used in a GLS regression to obtain the parameters of the "beta model", which are then used for simulation.

The second main difference, as noted above, involves the type of bootstrap method.[45] Version 2 of the Stata ELL package uses a traditional non-parametric bootstrap approach, which samples clusters randomly with replacement for each bootstrap replication. The R sae package takes a parametric bootstrap approach, which holds the composition of the sample constant for each replication. To do this, the method bootstraps the population instead of the sample. For each replication, a bootstrap population is constructed by drawing from the distribution of the estimated area random effect η and the household idiosyncratic error term ε. The EBP model is then re-estimated using the same households that were present in the sample, thus maintaining the same set of sample households for each replication. Values for η and ε are drawn 100 times from a normal distribution to generate point estimates for poverty in the population. The bootstrap process is repeated 100 times to estimate mean squared error, for a total of 10,000 simulations. The rationale for this parametric bootstrap procedure is that when performing small area estimation, the sample has been realized. The composition of the sample is therefore known and should be held constant across replications. This point is elaborated in greater detail in Diallo and Rao (2018).

A third difference between the EMDI and Stata SAE packages involves the model fitting method. The stata SAE package estimates uses either a traditional GLS model, or Henderson's method three to estimate the parameters of the random effects model. In the simulations below, we use Henderson's method three to fit the model in the SAE stata packages. The R EMDI package utilizes the maximum likelihood estimation method implemented in the lme function of the R nlme package (Pinheiro et al, 2020). In most settings, the choice of fitting method should not make a large difference to the model estimates.

A fourth difference is that the EMDI package, like the R SAE package, appends the sample to the census before aggregating poverty estimates. The Stata SAE package does not at this time. This should have a minor impact on the results in cases where the sample size is a small fraction of the census population.

A fifth difference is that empirical best estimation, along with the assumption of normal errors, is required in both version 3 of the Stata SAE package and the R EMDI package, while it is optional in version 2 of the Stata SAE package. The latter, accordingly, allows for non-parametric selection of the error terms when empirical best is not used, which relaxes the assumption of normality. We use version 2 of the Stata SAE package to estimate the model without empirical best estimation as a robustness check to verify that empirical best methods are much more efficient in this context.

[45] No software to our knowledge implements the Prasad and Rao (1990) approximation to obtain mean squared error estimates.