the review of income and wealth

Review of Income and Wealth
Series 68, Number S2, December 2022
DOI: 10.1111/roiw.12558

SMALL AREA ESTIMATION OF POVERTY UNDER STRUCTURAL CHANGE

BY SIMON LANGE

OECD Development Co-operation Directorate

UTZ JOHANN PAPE*

World Bank
University of Göttingen
AND

PETER PÜTZ

University of Göttingen

Small area estimation is an important tool to produce estimates of poverty for regions with low or zero sample sizes. Estimates are typically obtained by combining a consumption survey reporting on poverty and a census providing the spatial disaggregation. This paper discusses an updating method that produces up-to-date small area estimates when only a dated census and a more recent survey are available and predictors are subject to drift over time, a situation commonly encountered in practice. Instead of using survey variables to explain consumption in the survey, the updating approach uses only variables constructed from the census. The proposed estimator has fewer data requirements and weaker assumptions than common small area estimators. Applications to simulated data and to poverty estimation in Brazil show an overall good performance, but also imply the importance of examining practical challenges in real-world applications.

JEL Codes: D63, I32, R12

Keywords: poverty, population estimates, censuses, small area estimation

1. INTRODUCTION

Small area estimates are designed to produce precise estimates for regions with low or zero sample sizes. While small area estimates are useful for understanding poverty and inequality and for informing targeting decisions, household surveys are typically not designed to obtain reliable estimates at the level of cities, towns,

Note: The authors are listed in alphabetical order. The authors would like to thank Pierella Paci and Nobuo Yoshida as well as two anonymous referees for valuable comments, which greatly improved the paper. The authors also gratefully acknowledge feedback from participants of the fall 2019 IARIW-WB special conference, "New Approaches to Defining and Measuring Poverty in a Growing World," which was financially supported by the UK government through the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Programme. The views expressed are those of the authors and do not necessarily reflect those of the organizations they are affiliated with.

*Correspondence to: Utz Johann Pape, World Bank, Poverty and Equity Global Practice, 1818 H Street, NW Washington, DC 20433, USA (upape@worldbank.org).

districts, or villages. Censuses, on the contrary, do not provide information on consumption expenditures or income required to estimate the incidence of monetary poverty. To overcome these problems, poverty economists often use a method originally developed by Elbers *et al*. (2003) (henceforth, ELL) to combine census and survey data.

ELL's procedure consists of two steps. In the first step, a regression model is fitted with covariates that are recorded in both the survey and the census.[1] In the second step, the estimated model parameters are applied to census data. The regression model predicts the conditional mean of consumption. As one is typically also interested in at least the second moment of the distribution to obtain poverty estimates, the error distribution of the model is used to sample deviations from the conditional mean. The simulations provide estimates of consumption per capita for every household in the census. ELL show that estimators based on this procedure have levels of precision comparable to commonly used survey-based welfare estimators for populations as small as 15,000 households—roughly the size of a small town.

However, for this procedure to produce unbiased estimates of poverty measures at high levels of disaggregation, one key assumption must hold: The distribution of the explanatory variables is the same in both the census and the survey.[2] This assumption is problematic when some of the variables are subject to temporal changes and the census is not recent, that is, only a dated census and a more recent survey are available, a common situation in practice.[3] Reasons for a violation of this assumption include secular demographic trends, asset drift (Harttgen *et al*., 2013), and economic shocks.

In this paper, we show that reliance on cluster means over households from the census as predictors in the first step of the procedure yields poverty estimators that reflect the distribution of poverty rates at the time of the survey. We show theoretically and in simulations that this approach may still work adequately if the key assumption does not hold. In an application using Brazilian census data, we discuss practical challenges and provide guidance for using the proposed method.

The presented updating estimator is an unbiased estimator of poverty rates at the level of small areas if *aggregate* household characteristics from the dated census relate to consumption in the same way in clusters covered by the recent survey as in clusters not covered by the recent survey. When a survey used two-stage cluster sampling, the most common type of survey sampling, it will hold if clusters are indeed randomly selected.[4] When a master sampling frame with updated sampling weights is used for the recent survey, the estimation will still need to rely on the original sampling weights available for all clusters.

[1]The ELL estimator requires relevant explanatory variables for the model predicting consumption to be measured in a comparable way both in the census and in the survey, including the same degree of potential measurement error. Differences in coding schemes or even the way the interview was conducted can prevent reasonable harmonization between census and survey variables. See also Tarozzi and Deaton (2009).

[2]We also assume that real consumption is used, which is already appropriately spatially deflated.

[3]Censuses are usually conducted less frequently than surveys.

[4]Note also this assumption certainly must hold for the ELL method if the census and survey are conducted at the same time.

If the distribution of the explanatory variables remains stable over time, this procedure will typically result in less precision *vis-avis* ELL, with the magnitude of the loss depending on the difference in predictive power of the regression model in the first stage. We show that relying on the proposed approach might still yield useful results, but challenges in real-world small area applications must be properly accounted for.

ELL also suggests the additional use of census means over clusters to explain location effects, that is, cluster-specific effects. In this regard, the updating approach presented in this paper can be considered as a variant of ELL that does not use household-level variables. When we refer to the ELL method throughout this paper, what we have in mind is an estimator that combines survey and census variables at the household-level.

Cluster-averaged variables might also stem from different sources apart from census data. There is a growing literature showing that big data from mobile phones or remote sensing are readily available in many countries and predictive for up-to-date poverty measures at a high spatial resolution (e.g., Engstrom and Soundararajan, 2020; Koebe *et al*., 2021; Jean *et al*., 2016; Njuguna and McSharry, 2017; Pokhriyal and Jacques, 2017; Steele *et al*., 2017). As will be shown, the inclusion of such data in the methods discussed in this paper is straightforward.

There is by now a substantial literature about small area estimation especially in the context of poverty including alternatives and extensions to the ELL estimator. Tarozzi and Deaton (2009) and Molina and Rao (2010) argue that unexplained variation between areas impairs the performance of the ELL estimator as ELL only account for variation between clusters which are nested into areas. While also applying a two-stage approach similar to ELL, Molina and Rao (2010) use area-specific random effects instead of cluster-specific random effects. Moreover, in their empirical Bayes approach they simulate out-of-sample consumption values for the census conditional on the consumption values from the survey. Thus, in contrast to ELL, their simulation of the census data draws on observed sample information. Das and Chambers (2017) propose another correction for the ELL method which is robust to significant unexplained between-area variability. Their correction relies on the relationship between variance components estimators under the ELL model and a model that additionally contains an area-specific random effect. Marhuenda *et al*. (2017) discuss the direct application of such a model including cluster-specific and area-specific random effects for small area estimation via extending the empirical Bayes method of Molina and Rao (2010). Comprehensive discussion on different small area estimation methods can be found in Das and Haslett (2019), Guadarrama *et al*. (2016), and Haslett (2016). ELL is arguably the most frequently used small area estimation approach combining survey and census data.[5] Given its prominence, we compare our proposed approach with ELL.

There is also an emerging literature on updating small area estimates for poverty. However, to the best of our knowledge, the proposed updating procedure is the only one that does not rely on the key assumption of no drift in explanatory variables over time while not requiring any additional data. For instance, authors have

[5]According to Elbers and van der Weide (2014), it has been applied in more than 60 countries.

suggested relying only on explanatory variables assumed to be time-invariant (e.g. The National Statistical Coordination Board of the Philippines, 2021), while others have suggested "testing" for drift in explanatory variables (e.g. Ahmad *et al*., 2010). However, to the extent that there are empirical tests that support the assumption for a subset of predictors, severe shocks and extended time periods between survey and census will tend to quickly exhaust that subset, and it is exactly in those settings in which the demand for updated small area estimates is likely to be high.

Emwanu *et al*. (2006) require panel data with one wave collected at the time of the census. While structural changes in the explanatory variables may be detected and tackled by weighting procedures in such a setting, such procedures involve additional assumptions and uncertainties. Furthermore, the availability of panel data over a longer time span without substantial attrition is rare, especially in developing countries.

Isidro (2010) and Isidro *et al*. (2016) fit a model on simultaneously collected survey and census data first, for instance by ELL, and update the resulting estimates using a set of margins from a more recent survey. Thus, their approach requires contemporaneous survey and census collection with common variables as well as an up-to-date survey. As the method relies on updating multi-way contingency tables, it is computationally tractable only for a limited number of explanatory variables. A more general updating procedure is described in Betti *et al*. (2013). Their propensity score approach also aims at obtaining a covariate distribution in the census as if it was collected at the time of the recent survey. However, the method requires further modeling, including additional assumptions and uncertainty, and a survey collected at the time of the census.

Nguyen (2012) also discusses the updating method that we scrutinize in this paper. We complement Nguyen (2012), providing a more explicit discussion of the underlying assumptions and comparing the performance of the updating method to other estimators in a simulation setting and using real-world data with known ground truth. We also discuss in detail challenges faced when applying the updating estimator in real-world applications and provide guidance for these cases.

In the remainder of this paper, we show that the proposed updating method has comparably low data requirements and weak assumptions. Although our outcome variables will be measures of welfare, the method is applicable to a wide range of outcome measures and research questions beyond small area estimates for poverty.

The paper proceeds as follows: Section 2 presents the idea of the approach in detail. Section 3 describes the properties of the resulting poverty estimator. Simulation studies on artificial are presented in Section 4, while an application using real-world data from Brazil and inherent empirical challenges are discussed in Section 5, respectively. Section 6 concludes.

2. Estimating Poverty Measures Under Structural Change

Assume that the target population is a village v. The quantity of interest is a poverty measure W of the Foster-Greer-Thorbecke (FGT) family (Foster *et al*., 1984):

$$W_{\alpha v} = \frac{1}{N_v} \sum_{j=1}^{N_v} W_{\alpha vj} \tag{1}$$

with

$$W_{\alpha vj} = \left(\frac{z - y_{vj}}{z} \right)^{\alpha} I\left(y_{vj} < z\right), \qquad \alpha = 0,\ 1,\ 2.$$

Here, N_v is the size of the village population, y_{vj} is the consumption for individual j in village v, z is the poverty line, and $I(y_{vj} < z)$ is an indicator function which equals one if the consumption of an individual is below the poverty line and zero otherwise. Poverty headcount ratio, poverty gap, and poverty severity are obtained for $\alpha = 0$, 1, and 2, respectively.[6] Assume in the following that the aim is to estimate a poverty measure W, where the indices from (1) are dropped for notational convenience.

2.1. *The Consumption Model and Model Estimation*

In the following, we refer to consumption at the household level.[7] As most household consumption values are unobserved in a village, one needs a model that predicts those values for all households. Let y_{cht} be the consumption of household h in cluster c (e.g., an enumeration area) at time t. We consider the model

$$\begin{aligned} y_{cht} &= \mathbf{x}'_{c.,t-1}\boldsymbol{\gamma} + u_{ch} = \mathbf{x}'_{c.,t-1}\boldsymbol{\gamma} + \eta_{ct} + e_{cht}, \qquad h = 1, \cdots, H_c, \qquad c = 1, \cdots, C, \\ \eta_{ct} &\sim iid\ \mathcal{F}_1\left(0, \sigma_\eta^2\right), \qquad e_{cht} \sim iid\ \mathcal{F}_2\left(0, \sigma_e^2\right), \end{aligned} \tag{2}$$

which relates the (possibly transformed) consumption variable linearly to a vector $\mathbf{x}_{c.,t-1}$ containing dated census means of covariates over the cluster c from time point $t-1$.[8] The two error components are the cluster effects η_{ct} and the household errors e_{cht} which follow the (mutually independent) distributions $\mathcal{F}_1$ and $\mathcal{F}_2$, with mean zero and have variances σ_η^2 and σ_e^2, respectively. Heteroscedasticity can be modeled by relating the variance to a set of covariates. Such covariates may include the census means used in the main regression, but also higher moments such as the variance. Furthermore, geographic information and the fitted values of the first-stage regression may be used. The ELL method describes one option to model heteroscedasticity within the framework discussed here. Pinheiro and Bates (2000, ch. 5) provide a more comprehensive discussion.

In the first stage (2) is estimated using all household consumption values which are available for the village of interest in the survey. The estimation can be done by

[6]The proposed method is not restricted to measures of the FGT family, but applicable to essentially all measures which can be derived from consumption (or any other dependent variable measuring welfare), for instance inequality measures such as the Gini coefficient.

[7]Consumption expenditures are typically observed at the household level.

[8]In practice, one could use additional secondary information to explain consumption, for example, geographic information which is typically available in this context. Besides, fixed effects on higher aggregation levels such as counties and time-invariant explanatory variables on the household level could be, in principle, added to the consumption model. As discussed in Section 1, we do not assume many time-invariant variables to be available in practice and it is difficult to test if there are any. In this paper, we restrict ourselves to information that is available in the census.

weighted or (feasible) generalized least squares.[9] As the estimates are used to predict consumption values for the census, the aim is to find a model with high predictive power. Thus, one should find a parsimonious model containing only covariates that explain a substantial share of the variation in the dependent variable.

2.2. *Bootstrapping Census Consumption Data*

In the second stage, (2) is used to predict consumption values for each household in the village of interest based on the census. To be consistent with the first-stage model using the consumption values from the survey, the explanatory variables in the second stage are also averaged within clusters, that is, all households in the same cluster have the same value for each explanatory variable. In other words, the analysis is carried out at the cluster level. Using the estimated regression coefficients $\widehat{\boldsymbol{\beta}}$ from model (2) yields predictions $\widehat{y}_{cht} = \boldsymbol{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}}$, that is, predicted conditional means. To account for the deviations of the observed household consumption values from these means and the uncertainty involved in the estimation of $\boldsymbol{\beta}$, simulation techniques are used.

In particular, a bootstrap procedure is applied to generate R pseudo censuses and R resulting poverty measures estimates $\widehat{W}^{(r)}$:

1. Model coefficients are drawn from their respective sampling distribution estimated by the model in the first stage, including regression coefficients, random-term variances, and possible heteroscedasticity parameters.[10] For example, regression coefficients $\widehat{\boldsymbol{\beta}}^{(r)}$ are obtained for one sampling draw.
2. Conditional on the parameters describing the error components' distributions from the first step, cluster effects, and household errors are sampled from their respective distributions by a parametric or nonparametric bootstrap.[11] Independent of the strategy applied, bootstrapped error terms $\widehat{\eta}^{(r)}_{ct}$ and $\widehat{e}^{(r)}_{cht}$ are obtained.
3. Calculate the predicted consumption values $\widehat{y}^{(r)}_{cht} = \boldsymbol{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}}^{(r)} + \widehat{\eta}^{(r)}_{ct} + \widehat{e}^{(r)}_{cht}$ for all households as well as the poverty measure $\widehat{W}^{(r)}$ derived from those values.
4. Repeat steps 1–3 R times.

[9]The chosen estimation method depends on whether and how the survey design, potential heteroscedasticity, and the clustering nature of the data are considered.

[10]Usually, multivariate normal distributions with first-stage estimates for the means and the robust variance–covariance matrices accounting for correlation within clusters are used to draw the regression coefficients and potential heteroscedasticity parameters.

[11]For a parametric bootstrap, certain parametric distributions for the cluster effects and the household residuals must be determined. The sampling distributions for the nonparametric bootstrap are obtained directly from the first-stage estimates: a cluster effect can be estimated as the mean of the deviations between observed and predicted values in one cluster, that is, $\widehat{\eta}_{ct} = 1/H_c \sum_h^{H_c} (y_{cht} - \mathbf{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}})$, while the household residuals are computed as those deviations minus the cluster effects, that is, $\widehat{e}_{cht} = (y_{cht} - \mathbf{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}}) - \widehat{\eta}_{ct}$. There are different strategies to draw from these sampling distributions. One may draw with replacement from all estimated cluster effects and all household residuals. Alternatively, the household residuals may be drawn only from the location to which the drawn cluster effect belongs. This strategy generally allows the estimated two error components to be related in a nonlinear way, even though they are by construction uncorrelated.

For the poverty measure W, the (simulated) expected value is then given by

$$\tilde{\mu} = \frac{1}{R}\sum_{r=1}^{R} \widehat{W}^{(r)}, \tag{3}$$

and its variance by

$$\tilde{V} = \frac{1}{R}\sum_{r=1}^{R} (\widehat{W}^{(r)} - \tilde{\mu})^2. \tag{4}$$

Due to the bootstrap procedure, the variance contains uncertainty from the first-stage model (step 1, referred to as model error in the next section) and the unobservable part of consumption (step 2, referred to as idiosyncratic error in the next section).

3. Properties of the Estimator

We now turn to an investigation of the properties of the welfare estimator presented in the previous section. As described in ELL, the prediction error, the difference between the actual poverty measure W for a target population, say a village, and the proposed updating estimator $\tilde{\mu}$ of its expectation $E(W) = \mu$, is given by the sum of three terms:

$$W - \tilde{\mu} = (W - \mu) + \left(\mu - \hat{\mu}\right) + \left(\hat{\mu} - \tilde{\mu}\right), \tag{5}$$

where $\hat{\mu}$ is the expectation of $\tilde{\mu}$. If a sufficiently large number of bootstrap replications are conducted, the computation error, the third term in (5), can be ignored so that we will focus on the first and second terms.

The first term in (5), $(W - \mu)$, is the idiosyncratic error arising from the unexplained part of consumption of which the poverty measure is a function. Due to the stochastic nature of consumption, the actual poverty measure for a finite small area differs from expected poverty. As discussed in ELL, the idiosyncratic error vanishes asymptotically for growing population size, including additional clusters and individuals.

The second term, $\left(\mu - \hat{\mu}\right)$, is the model error, which originates from the estimation of (unknown) population parameters. The expectation of the model error equals zero if the poverty estimator is an unbiased estimator for the expected value of the true poverty measure. Whether this is the case depends on the regression model selected for the survey data.[12] What is crucial is that the distributional assumptions for the error components, namely the cluster effects and the

[12]It is neither intended nor necessary to establish causal or direct effects of explanatory variables on consumption. Thus, the regression coefficients in model (2) need not be unbiased or consistent with regard to the direct effects of the explanatory variables. In contrast, asymptotical unbiasedness of $\hat{\mu}$ can be obtained for several models, even if a single parameter in such a model might capture the effect of several correlated variables.

household errors, hold. Thus, one should check and potentially account for heteroscedasticity, serial correlation, and non-normality. The variance of the model error also depends fully on the properties of the first-stage estimators. It decreases in survey sample size.

If the assumptions of the ELL method hold and the models are correctly specified, the ELL estimator will usually exhibit a smaller variance of the prediction error than the updating estimator. The reason is that the latter is a between estimator that ignores variation within clusters. Intuitively, both estimators would only be similarly efficient if the explanatory variables differed distinctly more between clusters than within clusters. Another exception might occur in real-world applications if there are many missing values in the explanatory variables in the survey. Without imputation methods that are subject to estimation uncertainty, the ELL first-stage estimator would be based on a smaller sample than the updating estimator that uses census means.

In practice, the variance components of the idiosyncratic and model errors are not estimated separately. Rather, the entire variance of the prediction error is obtained from the variation of the simulated poverty estimates in equation (4). Therefore, under correct distributional assumptions on the random components, the bootstrap procedure allows to draw valid inferences, that is, to build confidence intervals that include the true poverty measure with a predetermined probability. For instance, bootstrap percentile intervals, which can be constructed directly from the bootstrap estimates (see Section 2.2), can be used for inference.

Another potential issue in practice is multicollinearity. Note that the fundamental unit of the predictors in the first stage is a cluster, not a household, and that the number of parameters that can be included in (2) is hence restricted to the number of clusters. However, household budget surveys that are used to estimate poverty incidence typically sample around 500 clusters with some sampling a substantially larger number. Therefore, we believe that the updating estimator could be based on a moderate number of regressors that would be sufficient to accurately predict household consumption which is assumed to differ between clusters.[13]

4. Simulation Experiments

A simulation study is conducted to compare the performance of the updating approach, ELL, and a purely survey-based estimator in predicting FGT poverty measures. We focus on the poverty headcount ratio W_0 and the poverty gap W_1 with three generic poverty lines that render 25 percent, 50 percent, and 75 percent of the population poor. The simulation setting is based on Tarozzi and Deaton (2009). In particular, the target population in the census is a single small area, say a town district with $N = 15{,}000$ households, divided into 150 clusters $k_c \in \{1, \cdots, 150\}$, each of size 100. In each simulation run, an artificial household survey is drawn from the census small area by selecting randomly 10 households from 100 randomly

[13]One commonly used rule-of-thumb is to restrict the number of predictors to the square root of observations. While our results in Sections 4 and 5 are based on fewer than 10 variables and 100 and 200 clusters, respectively, 500 clusters would allow the analyst to base the first-stage estimation on more than 20 census averages (or other summary statistics computed at the cluster-level).

selected clusters. First, both census and survey are generated by the following process with homoscedastic errors:

$$
\begin{aligned}
&y_{ch} = 25 + x_{ch} + \eta_c + e_{ch} \\
&x_{ch} = 0.01k_c - t_{ch}, \qquad k_c \in \{1, \cdots, 150\}, \qquad t_{ch} \sim U(0,1), \\
&\eta_c \sim N(0, 0.01), \qquad e_{ch} \sim N(0, \sqrt{2}).
\end{aligned}
$$

Thus, the explanatory variable is generated so that it differs in expectation between clusters. Such a situation with large and systematic differences in the averages of covariates across clusters (e.g., average levels of education or dwelling characteristics) is frequently observed in practice. This setting is ideal for the ELL method, which models this data-generating process. A linear regression based on the target population yields an R^2 of 0.55, while the updating method with an R^2 of 0.09 has considerably lower explanatory power.

A second setting mimics a real-word situation where the census is dated and a more recent household survey (with an underlying true census which is not observed) is available. Here the model explains consumption in the same way as the first setting for both the census and the survey, but the explanatory variable for the more recent survey is generated by

$$x_{ch} = 0.01k_c, \qquad k_c \in \{1, \cdots, 150\},$$

where the sampled 100 clusters in the survey have the same values for k_c as they have in the dated census. For both estimators, the R^2 obtained from the first-stage regression for all generated surveys is on average similar to the R^2 based on the census in the first setting.

In both settings, estimators purely based on the survey have desirable properties as the surveys are based on a random sample of decent size from the respective district population at the time of data collection. In real-word situations, however, a survey is not necessarily designed to obtain reliable estimates at the small area level, for example, as there are only few observations sampled.

All results are based on 300 Monte Carlo replications with 500 bootstrap census data sets generated in each replication for the two methods which use census data. The bootstrap procedure to sample the error components applies a simple nonparametric version, that is, both cluster effects and household errors are independently sampled with replacement from their sample analogs from the first-stage regression. See Section 2.2 for details.

In the first setting, the root mean squared error is, as expected, smallest for the ELL method, followed by the updating estimator and an estimator solely based on the survey (Table 1). Although the R^2 from the first-stage regression for the ELL method is seven times as large as for the updating method, the root mean squared errors only differ by a factor of about 1.5 or two-thirds, respectively. The coverage rates of the two methods are close to the nominal one of 95 percent and the bias is negligible.

In the second, more interesting setting, the ELL method naturally is the worst in terms of prediction and generates invalid confidence intervals (Table 2). The upward bias originates from the data-generating process above: as the expected values of x_{ch} and thus y_{ch} are larger in the recent survey and its underlying

TABLE 1
Monte Carlo Simulation Setting 1: Simultaneous Census and Survey Collection, and Some Variation in the Explanatory Variable Between Clusters

		Updating Estimator			ELL Estimator			Survey Est.
	True Value	Bias	RMSE	Coverage	Bias	RMSE	Coverage	RMSE
$W_0(.25)$	0.2500	0.0034	0.0122	0.9640	−0.0011	0.0084	0.9720	0.0139
$W_0(.50)$	0.5000	0.0099	0.0164	0.9480	0.0046	0.0104	0.9740	0.0183
$W_0(.75)$	0.7500	0.0076	0.0136	0.9360	0.0037	0.0093	0.9700	0.0159
$W_1(.25)$	0.0093	0.0002	0.0007	0.9380	−0.0001	0.0005	0.9700	0.0007
$W_1(.50)$	0.0242	0.0005	0.0012	0.9620	0.0000	0.0008	0.9700	0.0012
$W_1(.75)$	0.0477	0.0007	0.0016	0.9680	0.0001	0.0010	0.9780	0.0016

Notes: $W_\alpha(r)$ denotes the respective FGT measure for a poverty line that renders a share r of the population poor. The RMSE is the root of the mean squared deviations of the respective estimates from the true value over 300 replications. Coverage rates are calculated for 95% bootstrap percentile intervals.

TABLE 2
Monte Carlo Simulation Setting 2: Dated Census and Recent Survey, Some Variation in the Explanatory Variable Between Clusters, and Explanatory Variable Changes Over Time

		Updating Estimator			ELL Estimator			Survey Est.
	True Value	Bias	RMSE	Coverage	Bias	RMSE	Coverage	RMSE
$W_0(.25)$	0.2500	−0.0007	0.0117	0.9800	0.1375	0.1378	0.0000	0.0140
$W_0(.50)$	0.5000	0.0003	0.0134	0.9840	0.1053	0.1055	0.0000	0.0160
$W_0(.75)$	0.7500	−0.0031	0.0119	0.9640	0.0411	0.0416	0.0000	0.0151
$W_1(.25)$	0.0089	0.0000	0.0007	0.9440	0.0111	0.0112	0.0000	0.0007
$W_1(.50)$	0.0230	−0.0000	0.0011	0.9700	0.0156	0.0156	0.0000	0.0011
$W_1(.75)$	0.0454	−0.0001	0.0014	0.9860	0.0177	0.0177	0.0000	0.0014

Notes: $W_\alpha(r)$ denotes the respective FGT measure for a poverty line that renders a share r of the population poor. The RMSE is the root of the mean squared deviations of the respective estimates from the true value over 300 replications. Coverage rates are calculated for 95% bootstrap percentile intervals.

population than in the dated census, using the dated census data to predict current poverty statistics necessarily underestimates the current values of y_{ch} and hence overestimates poverty. In contrast, the updating method yields valid confidence intervals. It also results in a lower mean squared error in comparison to the purely survey-based estimate since additional census information is exploited. The last result typically holds on average if the model assumptions are fulfilled (as is the case in this simulation setting) and census and survey size differ distinctly. The latter is often true in practice.[14]

[14]Under the stated conditions, the updating estimator performs better only in predicting the true value on average. In a single sample, the pure survey mean is superior to the updating approach if the sample mean is by chance equal or very close to the census mean. An extreme example includes the limiting case in which the recent survey is equal to the underlying census. Then, the survey mean is trivially the census mean, that is, there is no error at all. However, the updating method is still prone to idiosyncratic and (small) simulation error, even under correct model specification.

5. Application to Brazilian Census Data

To evaluate and compare the proposed method in a real-world example and discuss practical challenges, we apply several poverty estimators to data from Brazil.

5.1. *Data and Model*

We use data extracts from the 2000 and 2010 Brazilian censuses provided by the Integrated Public Use Micro Sample (IPUMS, Minnesota Population Center, 2017), the preferred basis of welfare measurement in developing countries. Both censuses include information on monthly income at the level of the individual.[15] In addition, the data sets contain information that is potentially useful in explaining incomes, including the location in which the household resides (urban/rural), the number of household members, ownership of specific assets, as well as the employment status. This allows us to generate artificial surveys from the more recent census and predict income by dated census data. The poverty measures derived from the predicted income values can then be compared to actual poverty based on the recent census.

The census extracts comprise 2,652,356 and 2,906,184 observed households in 2000 and 2010 with full information on the used variables, respectively. The country is divided into 25 states and 1980 municipalities. As the municipalities constitute the smallest geographical unit that can be matched between 2000 and 2010, we consider the municipalities as clusters in the terminology used in the previous sections. We aim to estimate the household headcount ratio and the number of poor households at the state level, that is, the states are our small area of interest.[16] As most of the states are arguably larger than what would usually be considered a small area, we sample 5 percent of the observations of the census extracts. Furthermore, we drop five states consisting of fewer than 10 municipalities from our analyses as small areas typically include substantially more clusters.[17] The final sample sizes of the manipulated census extracts, which we will continue to refer to as "censuses" in the following, amount to 123,830 and 136,329 for the years 2000 and 2010, respectively. The obtained 20 small areas have realistic sample sizes and consist on average of a moderate number of municipalities (Table 3). To this end, we fit one model for all states and thereby ignore potential model heterogeneity between the states. We use averages over the municipalities for the 2000 census to predict household incomes in 2010. Household incomes are calculated as the sum of individual incomes of all household members, adjusted for the household size according to the OECD-modified scale.[18] To remove apparent right-skewness in the dependent variable, a log-transformation is applied after adding one to the household income values. The latter is done due

[15]The application here considers nominal incomes available from the Brazilian census data for the sake of demonstration and comparison. In practice, the approach should be implemented on the basis of real income or consumption expenditure, that is, income or consumption expenditure appropriately adjusted for spatial variation in prices.

[16]For the sake of illustration, we focus on calculating poverty estimates at the households level. Options for deriving poverty estimates at the individual level are discussed in Lange *et al*. (2018).

[17]Due to our sample scheme described below and to avoid unrealistically small clusters, we exclude municipalities with fewer than 25 households.

[18]http://www.oecd.org/eco/growth/OECD-Note-EquivalenceScales.pdf.

TABLE 3
Brazil 2010 Census Extract: Summary Statistics for the States

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
No. of households	1249.00	3048.50	4901.00	6816.45	8743.25	26,512.00
No. of municipalities	15.00	43.75	52.50	82.75	94.75	271.00
Poverty head count ratio	0.05	0.11	0.28	0.22	0.30	0.38

TABLE 4
Regression Results for the Updating Estimator Using All Households from Census of 2010

Dependent Variable: Income	
(Intercept)	− 0.69 (0.21)
Urban	− 0.06 (0.05)
Household size	− 0.11 (0.01)
Unemployed	− 2.14 (0.36)
Refrigerator availability	1.06 (0.05)
Computer availability	1.72 (0.20)
Phone availability	0.32 (0.06)
R^2	0.16
Adj. R^2	0.16
Num. obs.	136,329

Notes: The explanatory variables are averages over municipalities in the census of 2000. Standard errors clustered at the municipality level are shown in parentheses.

to the non-negligible amount of zero income values.[19] The poverty line is set to \$5.5 in 2011 PPP per person and day.[20]

As explanatory variables, we use the location (share of urban households), the average number of household members, the shares of households owning a phone, refrigerator and computer, respectively, as well as the unemployment rate. We additionally model heteroscedasticity using the same explanatory variables and by following the strategy explained in detail in ELL. When all households from the 2010 census are used, a linear regression with these explanatory variables yields an R^2 of 0.16 (Table 4). The estimated cluster effects variance in a linear mixed effects model based on the 2010 census is 0.02 and small compared to the estimated household residual variance of 0.94.

5.2. *Comparison Methods*

We compare the performance of the updating estimator for the headcount ratios of the 20 states with the ELL estimator and a simple (weighted) mean estimator based solely on the recent survey. While such a survey mean weighted for

[19]The proportion of all households in the 2010 census data with an income of zero amounts to 4.60 percent. As a robustness check, we also applied an inverse hyperbolic sine transformation to the dependent variable confirming the reported results.

[20]The set poverty line follows a recommendation given by the World Bank, see http://databank.worldbank.org/data/download/poverty/B2A3A7F5-706A-4522-AF99-5B1800FA3357/9FE8B43A-5EAE-4F36-8838-E9F58200CF49/60C691C8-EAD0-47BE-9C8A-B56D672A29F7/Global_POV_SP_CPB_BRA.pdf.

the sampling frame is an unbiased poverty estimator by construction, it may come along with unacceptably high variance because only small sample sizes are available for single small areas. Even no observation might be sampled from a single small areas. To this end, the presented setting emulates a real-world application in the small area context and is thus appropriate for using more advanced estimators such as the ELL or the updating approach.

For the ELL first-stage regression, the same aggregated explanatory variables as for the updating approach are used, but also the respective non-aggregated household level variables from the recent survey. In a regression based on all households from the 2010 census, this simple model specification yields an R^2 of 0.35.

When predicting the total number of poor households for the states, we compare the updating method to a further estimator that is motivated in the following section.

5.3. *Discussion of Practical Challenges*

Mimicking a real-world, two-stage sampling strategy, we draw 300 artificial surveys from the 2010 census by first sampling 200 municipalities with a probability proportional to their size in the 2000 census as the municipality sizes are known only at this time.[21] At the second stage, we sample 25 households randomly from each of those municipalities, resulting in an overall survey sample size of 5000 households. If the number of households in some municipalities grows faster or slower over time than in others, the model estimation at the first stage must account for these differences by using appropriate weights. This requires knowledge of the number of households in the municipalities at the time of the survey. We assume the number of households in the sampled municipalities to be known from a listing exercise, but to be unknown in the remaining municipalities. The latter may bias all estimators that include estimates or sample sizes from non-sampled municipalities. For computing the headcount ratio, the survey-based estimator does not rely on any information from non-sampled municipalities. However, when estimating the total number of poor in a state, the number of households in the small area at the time of the survey needs to be known also for this estimator. The ELL method and the updating approach are constructed as composite measures of all municipalities, that is, they rely on sampled and non-sampled municipalities. A bias occurs if the number of households in some municipalities grows faster or slower over time than in others and if this unequal growth is related to the poverty indicator of interest. In our data, the number of households per municipality grows on average by 38 percent over the 10 years under consideration. However, there is considerable variation in growth rates, which range from −53 percent to 309 percent. The corresponding change over time is only weakly correlated with the true headcount ratios ($r = 0.08$) as well as with the true number of poor households in the municipalities ($r = -0.06$) when considering all states jointly. In single states, correlations within the range of $[-0.13;0.37]$ and $[-0.46;0.12]$ occur for the headcount ratio and the

[21]While all municipalities are measured in both censuses and thus all sampled municipalities are available both in the artificial surveys and the outdated census, in real-world applications the number of overlapping clusters between the survey and the census can be small.

number of poor households, respectively. Thus, the resulting bias for estimators relying on non-sampled municipalities is considerable in single small areas and might be even also large in other settings. In practice, expert knowledge in combination with the insights from the observed clusters may help to gauge the severity of this issue, but especially for single small areas with few sampled clusters the uncertainty might be huge. In this application, we predict for each non-sampled municipality the number of households in 2010 by adjusting the respective number of households in 2000 by the average growth for the sampled municipalities, that is, each non-sampled municipality is assumed to grow by the same rate.[22]

For the described two-stage estimators, we use a weighted linear regression in the first stage that accounts for different changes in population sizes between clusters over time. In the second-stage bootstrap procedure, the regression coefficients are sampled from a multivariate normal distribution where the expected values and the cluster-robust variance-covariance matrix are estimates from the first stage. The error components are generated by a nonparametric bootstrap. In particular, cluster effects are drawn with replacement from the 200 first-stage estimates. The household errors are drawn with replacement from the first-stage residuals belonging to this specific cluster. See also Section 2.2. We set the number of bootstrap replicates applied in the second stage to 150.

If we are interested in the total number of poor households, a further regression estimator that does not rely on unknown municipality sizes can be used. We will call this estimator the aggregation estimator. The idea is to use as well dated census means at the municipality level as explanatory variables, but to use only one observation per municipality and predict the total number of poor households in the municipality. To this end, one first estimates a weighted regression model that accounts for the probability proportional to size sampling, that is, setting weights inverse to the municipality size in 2000. The dependent variable is the number of observed poor households per municipality projected for the number of households in the respective municipality. We use the same explanatory variables as described above in a linear regression model, yielding an R^2 of 0.09 when using all 1655 municipalities.[23]

5.4. *Results*

Considering the headcount ratio estimates for 20 states, the updating estimator exhibits a non-negligible bias for single states that may be due to the mainly unknown municipality sizes or unmodeled heterogeneity between states as described above (Table 5). Accordingly, the coverage of the bootstrap confidence intervals is sometimes far below the nominal one of 95 percent (Figure 1). As expected, the survey estimates on average approach the true values for the vast majority of states, while the ELL estimator is severely biased due to changes in the

[22]One could also apply a bootstrap approach, that is, sampling with replacement from the observed municipality growth rates and apply them to the non-sampled municipalities. We refrain from this strategy here as it would be computationally intensive and introduce more uncertainty to the results.

[23]We tried different specifications for this method such as transforming the dependent variable and using counts models but they all yielded worse prediction accuracy.

TABLE 5
HEADCOUNT RATIO AT STATE LEVEL

Measure	Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Bias	Updating est.	−0.0497	−0.0260	0.0108	0.0028	0.0268	0.0461
	ELL	0.0292	0.0535	0.0667	0.0700	0.0892	0.1168
	Survey est.	−0.0337	−0.0060	−0.0008	−0.0034	0.0017	0.0171
CV	Updating est.	0.0387	0.0531	0.0570	0.0597	0.0649	0.0920
	ELL	0.0335	0.0385	0.0425	0.0446	0.0499	0.0691
	Survey est.	0.1157	0.1562	0.2055	0.2306	0.2683	0.4338
MAE	Updating est.	0.0080	0.0208	0.0265	0.0274	0.0353	0.0500
	ELL	0.0290	0.0532	0.0670	0.0701	0.0892	0.1170
	Survey est.	0.0120	0.0177	0.0350	0.0388	0.0490	0.1080
RMSE	Updating est.	0.0100	0.0240	0.0300	0.0300	0.0372	0.0520
	ELL	0.0320	0.0545	0.0675	0.0712	0.0902	0.1180
	Survey est.	0.0150	0.0225	0.0440	0.0493	0.0635	0.1280

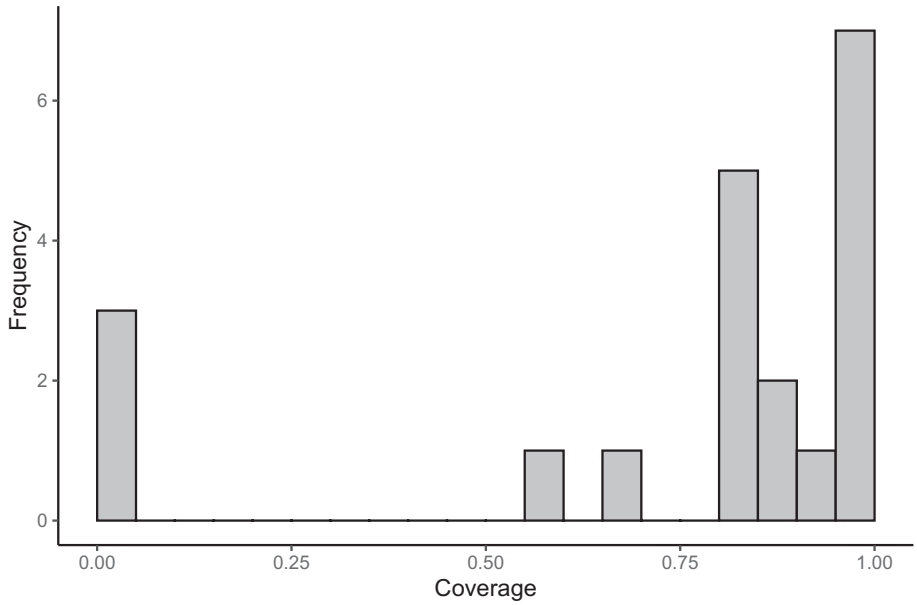

Figure 1. Distribution of Coverage Rates Over States
Note: A coverage rate for a state is the fraction of 300 Monte Carlo simulations for which the true value of the poverty headcount ratio lies within the 95 percent percentile interval computed on 150 bootstrap replication.

distribution of the explanatory variables over time (Table 6).[24] For instance, the share of households owning a computer increased from 8 to 31 percent between 2000 and 2010. The coefficient of variation is much smaller for the two small area

[24]The survey estimator yields small biases for some states. In these states, often only one municipality is sampled, for example, due to the presence of one very large and otherwise very small municipalities. However, correct adjustment for changing municipality sizes needs at least two sampled municipalities.

TABLE 6
Comparison of Explanatory Variables Over Time

Year	2000	2010
Urban	0.79	0.79
Household size	3.78	3.29
Unemployed	0.06	0.07
Refrigerator availability	0.79	0.90
Computer availability	0.08	0.31
Phone availability	0.34	0.31
N	123,830	136,329

TABLE 7
Number of Poor Households at State Level

Measure	Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Bias	Updating est.	−253.45	−129.84	−5.93	11.44	115.78	414.12
	ELL	−0.26	118.39	301.50	360.76	415.25	1271.41
	Survey est.	−60.39	−16.12	18.43	53.59	76.50	307.25
	Aggregation est.	−655.66	−161.07	−44.60	−39.44	81.12	702.96
CV	Updating est.	0.04	0.06	0.06	0.07	0.07	0.09
	ELL	0.04	0.05	0.05	0.05	0.06	0.07
	Survey est.	0.41	0.43	0.45	0.46	0.47	0.57
	Aggregation est.	0.06	0.09	0.10	0.11	0.12	0.20
MAE	Updating est.	20.91	86.53	143.61	162.12	222.12	414.12
	ELL	18.61	118.39	301.50	364.54	415.25	1271.41
	Survey est.	89.74	155.53	220.80	228.58	261.22	455.28
	Aggregation est.	36.73	90.76	146.94	217.70	255.90	702.96
RMSE	Updating est.	26.11	94.54	163.31	175.18	228.54	431.49
	ELL	23.56	121.05	310.00	372.89	422.01	1279.81
	Survey est.	119.17	199.72	285.43	290.89	329.71	578.50
	Aggregation est.	44.06	110.24	171.04	240.02	305.89	727.47

methods that exploit census information. The measures for predictive accuracy, namely the mean absolute error and the root mean squared error, are arguably most relevant from a practitioner's perspective. Due to the large bias of the ELL method and the high variability in the survey-based estimator, the updating method performs best with regard to both of these measures. The survey-based method has high prediction errors for single states. Furthermore, in 27 simulation runs there is at least one state with no observation sampled such that no poverty estimate can be calculated using this method.

The results are qualitatively similar if the number of poor households is the quantity of interest (Table 7). With regard to the mean absolute error and the root mean squared error, the additional comparison method using one observation per municipality, the "aggregation estimator," performs slightly better than the survey-based estimator, but worse than the updating method. However, in other applications additional bias might worsen the prediction accuracy of the updating method. Put differently, in such cases the small area updating method including several modeling choices might not necessarily outperform a much simpler estimator.

6. Conclusions

In this paper we examined a common situation in small area estimation for poverty, namely the availability of a recent survey and dated census data. We presented an approach with low data requirements and weak assumptions for this situation that showed an overall good performance. If the distribution of explanatory variables changes over time, this updating estimator is superior to the most frequently used methods for contemporaneous census and survey collection. It may also provide more precise area-specific estimates than a direct estimate based solely on the survey.

However, the proposed approach is not immune to systematic migration patterns or demographic developments over time and thus involves additional assumptions and uncertainties. Furthermore, it is more difficult to obtain a sound model with sufficient explanatory power that holds for all small areas of interest when using old census data to predict recent income values. Including data from other sources, for example, from mobile phones or satellite imagery, may be helpful in this regard.

Likewise, issues typically encountered in small area estimation techniques that combine census and survey data must be considered. In particular, variable selection and adequate modeling of apparent heteroscedasticity and differences in skewness in the error term can be challenging. Violations of the assumptions on the error term may be partly solved by allowing for more distributional flexibility in the response variable or the error term. Rojas-Perilla *et al*. (2020) and the references therein provide various transformations of the response variable to achieve the validity of the assumption of identically and normally distributed error terms. A more comprehensive approach would be the application of Generalized Additive Models for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005). This framework not only includes a huge variety of potential response distributions, but also allows to link all parameters of those distributions to explanatory variables. This allows for a straightforward way to model heteroscedasticity and skewness simultaneously in one coherent model. Moreover, nonlinear and spatial effects can be integrated into the GAMLSS framework. Although model choice is also a challenging task, it might be a very interesting direction for future research to combine GAMLSS and existing small area approaches, irrespective of the time span between census and survey collection.

In practice, all these empirical challenges may lead to estimates that are no more useful than those from much simpler estimators. If careful model building and checking do not help, it is time to conduct a new census.

References

Ahmad, N., F. Ahmed, D. M. Jolliffe, M. A. Khan, M. A. Mahbub, l Sharif, V. Swaroop, N. Yoshida, S. Zaidi, and J. C. Zutt, "Poverty Maps of Bangladesh 2010," 2010. Available from http://documents.worldbank.org/curated/en/160611468014459434/Technical-report, accessed on 10 September 2021.

Betti, G., A. Dabalen, C. Ferré, and L. Neri, "Updating Poverty Maps between Censuses: A Case Study of Albania," *Poverty and Exclusion in the Western Balkans*. Springer, 2013.

Das, S., and R. 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, 1137–61, 2017.

Das, S., and S. Haslett, "A Comparison of Methods for Poverty Estimation in Developing Countries," *International Statistical Review*, 87, 368–92, 2019.
Elbers, C., and R. van der Weide, "Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality." World Bank Policy Research Working Paper No. 6962, The World Bank, 2014.
Elbers, C., J. O. Lanjouw, and P. Lanjouw, "Micro-Level Estimation of Poverty and Inequality," *Econometrica*, 71, 355–64, 2003.
Emwanu, T., J. G. Hoogeveen, and P. Okiira Okwi, "Updating Poverty Maps with Panel Data," *World Development*, 34, 2076–88, 2006.
Engstrom, R., D. Newhouse, and V. Soundararajan, "Estimating Small-area Population Density in Sri Lanka Using Surveys and Geo-spatial Data," *PLoS ONE*, 15, e0237063, 2020.
Foster, J., J. Greer, and E. Thorbecke, "A Class of Decomposable Poverty Measures," *Econometrica*, 52, 761–66, 1984.
Guadarrama, M., I. Molina, and J. N. K. Rao, "A Comparison of Small Area Estimation Methods for Poverty Mapping," *Statistics in Transition New Series*, 1, 41–66, 2016.
Harttgen, K., S. Klasen, and S. Vollmer, "An African Growth Miracle? Or: What Do Asset Indices Tell Us About Trends in Economic Performance?" *Review of Income and Wealth*, 59, 761–66, 2013.
Haslett, S. J., "Small Area Estimation Using Both Survey and Census Unit Record Data." *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons, Ltd, 2016.
Isidro, M. C., "Intercensal Updating of Small Area Estimates: A Thesis Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Statistics at Massey University, Palmerston North, New Zealand," 2010. PhD thesis, Massey University, New Zealand.
Isidro, M. C., S. Haslett, and G. Jones, "Extended Structure Preserving Estimation (ESPREE) for Updating Small Area Estimates of Poverty," *Annals of Applied Statistics*, 10, 451–76, 2016.
Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining Satellite Imagery and Machine Learning to Predict Poverty," *Science*, 353, 790–94, 2016.
Koebe, T., A. Arias-Salazar, N. Rojas-Perilla, and T. Schmid, "Intercensal Updating Using Structure-preserving Methods and Satellite Imagery," arXiv preprint 2103.03834, 2021.
Lange, S., U. J. Pape, and P. Pütz, "Small Area Estimation of Poverty under Structural Change," World Bank Policy Research Working Paper No. 8472, The World Bank, 2018.
Marhuenda, Y., I. Molina, D. Morales, and J. N. Rao, "Poverty Mapping in Small Areas Under a Twofold Nested Error Regression Model," *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 180, 1111–36, 2017.
Minnesota Population Center, "Integrated Public Use Microdata Series," International: Version 6.5, 2017.
Molina, I., and J. N. Rao, "Small Area Estimation of Poverty Indicators," *Canadian Journal of Statistics*, 38, 369–85, 2010.
Nguyen, V. C., "A Method to Update Poverty Maps," *Journal of Development Studies*, 48, 1844–63, 2012.
Njuguna, C., and P. McSharry, "Constructing Spatiotemporal Poverty Indices from Big Data," *Journal of Business Research*, 70, 318–27, 2017.
Pinheiro, J. C., and D. M. Bates, *Mixed Effects Models in S and S-PLUS*. Springer, 2000.
Pokhriyal, N., and D.C. Jacques, "Combining Disparate Data Sources for Improved Poverty Prediction and Mapping," *Proceedings of the National Academy of Sciences*, 114, E9783–92, 2017.
Rigby, R. A., and D. M. Stasinopoulos, "Generalized Additive Models for Location, Scale and Shape," *Applied Statistics*, 54, 507–54, 2005.
Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis, "Data-Driven Transformations in Small Area Estimation," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183, 121–48, 2020.
Steele, J. E., P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y. A. de Montjoye, A. M. Iqbal, et al., "Mapping Poverty Using Mobile Phone and Satellite Data," *Journal of The Royal Society Interface*, 14, 20160690, 2017.
Tarozzi, A., and A. Deaton, "Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas," *Review of Economics and Statistics*, 91, 773–92, 2009.
The National Statistical Coordination Board of the Philippines, "2003 City and Municipal Level Poverty Estimates," 2009. Available from https://psa.gov.ph/content/city-and-municipal-level-poverty-estimates accessed on 10 September 2021.