WPS8472

Policy Research Working Paper 8472

Small Area Estimation of Poverty under Structural Change

Simon Lange
Utz Johann Pape
Peter Pütz

WORLD BANK GROUP
Poverty and Equity Global Practice
June 2018

Policy Research Working Paper 8472

Abstract

Small area poverty maps allow for the design of policies based on spatial differences in welfare. They are typically estimated based on a consumption survey reporting on poverty and a census providing the spatial disaggregation. This paper presents a new method which allows for the estimation of up-to-date small area poverty maps when only a dated census and a more recent survey are available and predictors and structural parameters are subject to drift over time, a situation commonly encountered in practice. Instead of using survey variables to explain consumption in the survey, the new approach uses variables constructed from the census. The proposed estimator has fewer data requirements and weaker assumptions than common small area poverty map estimators. Applications to simulated data and to poverty estimation in Brazil show an overall good performance.

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/research. The authors may be contacted at upape@worldbank.org.

Small Area Estimation of Poverty under Structural Change[1]

Simon Lange[2], Utz Johann Pape[3], Peter Pütz[4]

Keywords: Poverty, Population estimates, Censuses, Small Area Estimation

JEL classification: D63, I32, R12

[1]Authors in alphabetical order. Findings, interpretations and conclusions expressed in this paper are entirely those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments of the countries they represent. The authors would like to thank Pierella Paci and Nobuo Yoshida for valuable comments on earlier drafts.

[2]World Bank, Poverty and Equity Global Practice, Africa.

[3]World Bank, Poverty and Equity Global Practice, Africa. Corresponding author. E-mail: upape@worldbank.org.

[4]Economics Department, University of Göttingen.

1 Introduction

A poverty map is a spatial description of the distribution of poverty in a given country or region. While such a map is useful for policy makers and researchers when small geographic units (e.g., cities, towns, or villages) are discernable, estimates based on household surveys are typically not representative or associated with high uncertainty at such levels of disaggregation. On the other hand, most censuses do not contain information on consumption (or a surrogate such as income or expenditures) required to calculate poverty. To overcome these problems, Elbers et al. (2003), henceforth ELL, developed small area estimation poverty maps, a methodology that can be used to combine information from a detailed household survey with that from a comprehensive census. The general methodology usually consists of two steps, calibration of a statistical model based on survey data and application to the comprehensive census data. In the first step, a multiple linear regression analysis is used to estimate a model of household consumption based on survey data (which includes a consumption module). The explanatory variables in the model are restricted to the subset available 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 Since one is typically also interested in higher moments of the distribution, simulation methods are used to introduce a random disturbance term. The simulations provide estimates of consumption per capita for every household in the census.

Several criticisms have been raised with regard to the ELL estimator and extensions and alternatives have been discussed. Haslett et al. (2010) propose alternative regression techniques to estimate the survey regression in the first stage. 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 simulated census data explicitly includes 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 which additionally contains an area-specific random effect. Marhuenda et al. (2017) discusses the direct application of such a model including cluster-specific and area-specific random effects for poverty mapping via extending the empirical Bayes method of Molina and Rao (2010). Comprehensive discussions on different small area estimation methods can be found in Guadarrama et al. (2016) and

[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) for a brief discussion.

Haslett (2016). Still, ELL's is arguably the most frequently used poverty mapping approach combining survey and census data. According to Elbers and van der Weide (2014), it has been applied in more than 60 countries. Some examples for the application of ELL, including in areas other than poverty mapping, are Healy et al. (2003), Demombynes and Özler (2005), Elbers et al. (2007), Araujo et al. (2008), Agostini et al. (2010), Bui and Nguyen (2017) and Gibson (2018).

A key assumption for the applicability of ELL is that the distribution of the explanatory variables is the same in both census and survey. This assumption will often be violated if time has passed between data collection for the census and survey, i.e., only a dated census and a more recent survey are available, a common situation as censuses are usually conducted less frequently than surveys. Reasons for a violation of this assumption may include demographic trends, migration, natural disasters, and conflicts. If the population parameters, including the regression coefficients, remain unchanged but the distributions of the explanatory variables change over time, ELL results in an outdated poverty map, namely a poverty map at the time of the census. If both the population parameters and the explanatory variables change over time, it is not quite clear what is obtained, but generally not an up-to-date poverty map.

The discussed assumptions on the explanatory variables can be relaxed if household characteristics from the census are used to explain consumption values from the survey in the first stage to obtain parameter estimates. These can then be used to predict consumption values using the census data in the second stage. As it is usually impossible to match households between a census and a survey, the estimation needs to be conducted at a higher geographical level, for instance at the level of census enumeration areas. Throughout this paper, we will refer to the generic term of clusters as the lowest level at which census and survey information can be matched. If the assumptions on the explanatory variables hold, this aggregation may worsen the prediction accuracy vis-à-vis ELL, with the magnitude of the loss of precision hinging on the regression model in the first stage. Note that ELL also propose the additional use of census means to explain location effects, i.e. cluster-specific effects. In this regard, our approach can be considered as a variant of ELL without the use of household-level variables included in both census and survey and without reliance on the associated assumptions. When we refer to the ELL method throughout this paper, we have in mind an estimator that combines survey and census variables at the household-level, the central idea of the approach.

In the case that at least one of the underlying assumptions of ELL is violated, our new approach will still produce up-to-date poverty maps with unbiased poverty estimates. The key assumption we introduce is that *aggregate* household characteristics from the old census relate to consumption the same way in clusters covered by the new survey as in clusters not covered by the new survey. This assumption will hold (on average) if clusters are randomly drawn. Note that a similarly weak assumption has to be made for the applicability of the ELL method if the census and survey are conducted at the same time, namely

that household characteristics from the survey relate to consumption the same way in clusters covered by the survey as in clusters not covered by the survey.

In a different scenario, a recent census and only dated survey data may be available. Reliable predictions of poverty measures at the time of the recent census can only be obtained under the additional strong assumption of non-changing structural parameters (including the regression parameters linking explanatory variables to consumption) over time (e.g., Kijima and Lanjouw, 2003). This holds for both ELL and our estimator. If both structural parameters and the distribution of the explanatory variables change over time, ELL results in biased estimates. In contrast, linking census covariate means to predict survey consumption would remain a valid method to generate a poverty map at the time of the survey. In the remainder of this paper, we will focus on the practically more relevant case of a dated census and a recent survey.

Although monitoring poverty over time is of eminent interest to economists (see, for instance, Deaton and Kozel, 2005), little attention has been paid to updating small area estimation approaches which combine dated census and recent survey data. 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, the remaining assumptions of the ELL method as described above are still required. Furthermore, availability of panel data over a longer time span without substantial attrition is rare, especially in developing countries. The National Statistical Coordination Board of the Philippines (2009) uses explanatory variables deemed time-invariant to estimate intercensal poverty measures. Whether the distribution of variables changes over time is not assessed formally but rather based on *impromptu* assumption. This approach still relies on similar assumptions as the ELL method, even though changes in the distribution of the explanatory variables are ruled out by choosing time-invariant variables. One may also test whether the distribution of potential predictors changed over time and then restrict the set of predictors in the first stage to only those that exhibit no drift.[2] However, severe shocks and extended time periods between survey and census will tend to quickly exhaust the set of viable predictors to do so. And it is exactly in those settings in which the demand for an updated poverty map is likely to be high. Isidro (2010) and Isidro et al. (2016) propose to fit a model on simultaneously collected survey and census data first, for instance by ELL, and update the resulting estimates using a more recent survey. Their Extended Structure Preserving Estimation (ESPREE) approach does not require panel data but contemporaneous surveys and census collection with common variables. The ESPREE method relies on updating multi-way contingency tables which is computationally tractable only for a limited number of categorical explanatory variables and an outcome indicator which is a proportion, for instance the number of people who live below the poverty line. A more general updating

[2]This has been suggested for an update of the Bangladeshi poverty maps by researchers from The Bangladesh Bureau of Statistics, The World Bank and The United Nations World Food Programme (2010).

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 modelling, including additional assumptions and uncertainty, and a survey collected at the time of the census.

In the remainder of this paper, we show that our proposed method has comparably low data requirements and weak assumptions. Although our outcome variables will be measures of welfare, our method is applicable to a wide range of outcome measures and research questions beyond poverty mapping. Section 2 presents the idea of the approach in detail. Section 3 describes the properties of the resulting poverty estimator. Simulation studies on artificial and real data are presented in Sections 4 and 5, respectively. Section 6 concludes.

2 Estimating poverty measures under structural change

Assume that the target population is a village v. While the proposed method is 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, assume for now that the measures of interest are poverty measures 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

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

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.

2.1 The consumption model

Usually, consumption values are observed at the level of the household, not the level of the individual. As most household consumption values are unobserved in a village, one needs a model which predicts those values for all households. Let y_{cht} be the consumption of household h in cluster c at time t. Then, the model of consideration is

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

which relates the (potentially 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 $t-1$.[3] The two error components are the cluster effects η_{ct} and the household errors e_{cht} which follow the distributions $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively, and are assumed to be independent of each other. It is possible to allow for heteroscedasticity in the household error by modeling its variance to 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, while Pinheiro and Bates (2000, ch. 5) provide a more comprehensive discussion.

2.2 Model estimation based on survey consumption values

In the first stage, model (2) is estimated using all household consumption values which are available for the village of interest in the survey. The estimation can be done within the maximum likelihood framework or by weighted or (feasible) generalized least squares.[4] 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 which explain a substantial share of the variation in the dependent variable. Due to averaging over the cluster, means over candidate variables should exhibit variation across clusters.

2.3 Bootstrapping census consumption data

In the second stage, model (2) is used to predict consumption values for each household in the village of interest based on the census. Note that, 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, i.e., all households in the same cluster have the same value for each explanatory variable. Using the estimated regression coefficients $\widehat{\boldsymbol{\beta}}$ from model (2) yields predictions $\hat{y}_{cht} = \boldsymbol{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}}$, i.e. predicted

[3] In practice, one could use additional secondary information to explain consumption, e.g. geographic information which is typically available in poverty mapping exercises. Besides, fixed effects on higher aggregation levels such as counties and time-invariant explanatory variables on the household level $\mathbf{x}_{cht}$ 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.

[4] The chosen estimation method depends on whether and how the survey design, potential heteroscedasticity and the clustering nature of the data are taken into account.

conditional means. To account for the deviations of the observed consumption values from these means, random disturbance terms have to be added by simulation. Assume that the aim is to estimate a poverty measure W, where the indices from (1) are dropped for notational convenience.

A bootstrap procedure is applied to generate R pseudo censuses and resultant poverty measures:

1. Draw all model coefficients from their respective sampling distribution estimated by the model in the first stage, including regression coefficients, random term variances and possible heteroscedasticity parameters. Multivariate normal distributions with first-stage estimates for the means and variance-covariance matrix are used to draw the regression coefficients and the heteroscedasticity parameters.[5]

2. Conditional on the parameters describing the error components' distributions from the first step, cluster effects and household errors are drawn from their respective distributions. One option is to use a parametric bootstrap, i.e., to assume certain parametric distributions for which the estimates from the first stage regression might give some indication. However, a nonparametric bootstrap procedure is a valid alternative or supplement. In this case, a cluster effect can be estimated as the mean of the deviations between observed and predicted values in one cluster, i.e. $\widehat{\eta}_{ct} = 1/H_c \sum_h^{H_c} (\widehat{y}_{cht} - \mathbf{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}})$, while the household residuals are computed as those deviations minus the cluster effects, i.e., $\widehat{e}_{cht} = (\widehat{y}_{cht} - \mathbf{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}}) - \widehat{\eta}_{ct}$. 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 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.

3. Calculate the predicted consumption values for all households and all individuals as well as the poverty measure $\widehat{W}^{(r)}$ derived from those values.

4. Repeat steps 1 to 3 R times.

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

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

[5] One may also assume a distribution for the error components' variances such as the gamma distribution, e.g., but in many cases it is reasonable to treat their estimates from the first stage as fixed, especially if the numbers of enumeration areas and households in the survey are large since then there is not much uncertainty in the variance estimators. The household error variance estimator is usually very precise as it is based on the (large) number of households in the survey. The amount of enumeration areas in the survey is smaller but the uncertainty in the variance estimator of the enumeration area effects is often still negligible. In practice, one may check whether the estimated variances of the error components' variances are small enough in order to treat them as fixed in all bootstrap replications.

and its variance by

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

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

In the following, we will investigate the properties of our welfare estimator presented in the previous section.

As described in ELL, the prediction error, the difference between the true poverty measure W for a target population, say a village, and our estimator $\widetilde{\mu}$ of its expectation $E(W) = \mu$, is given by

$$W - \widetilde{\mu} = (W - \mu) + (\mu - \widehat{\mu}) + (\widehat{\mu} - \widetilde{\mu}). \tag{5}$$

Here, the third component is the computation error which is the difference between our estimator $\widetilde{\mu}$ and its expectation $\widehat{\mu}$. In the following, we assume the computation error to be negligible by applying a sufficiently high number of bootstrap simulations.

The first term on the right-hand side of equation (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 true poverty measure differs from its expected one. Note that the population in the small area of interest is finite and can be seen as a realization from an infinite population. Hence, all asymptotic results for the idiosyncratic error of the poverty measure from ELL carry over to the new approach presented here: the idiosyncratic error vanishes asymptotically for growing population size, including additional clusters and individuals.

The second part of equation (5), $(\mu - \widehat{\mu})$, 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 hinges on the regression model selected for the survey data.[6] What is crucial is that the assumptions of zero mean, independence, and homoscedasticity for the error components, namely the cluster effects and the household errors, hold. Likewise, if the error components are assumed to follow certain distributions

[6] Note that 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 estimated unbiasedly or consistently with regard to the direct effects of the explanatory variables. In contrast, asymptotical unbiasedness of $\widehat{\mu}$ can be obtained for several models, even if a single parameter in such a model might capture the effect of several correlated variables.

and these parametric assumptions are used for the generation of simulated census data sets (see Section 2.3), they also have to hold. Note that these assumptions may be valid even if dated census data are used for predicting survey consumption values. Thus, one crucial part is the diagnosis of the estimated error components from the first-stage regression. If plots or statistical tests on the estimated cluster effects and residuals suggest violations of distributional assumptions, one should adjust the model accordingly. More specifically, heteroscedasticity, serial correlation, and non-normality can be detected and accounted for, for instance by choosing different predictor specifications, transforming the dependent variable, or explicit modeling of heteroscedasticity as discussed in Section 2.1. 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 our 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. In practice, another exception might occur 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 our estimator.

In practice, the variance components of the idiosyncratic and model error 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). Hence, under correct distributional assumptions on the random components, the bootstrap procedure allows to draw valid inferences, i.e., to build confidence intervals which 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 2.3), 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 cover 500 clusters or more, with some covering substantially more. Hence, we believe that our 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.[7]

[7] 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 100 clusters and less than ten variables, 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).

4 Simulation experiments

A simulation study is conducted to compare the performance of our approach, ELL, and a purely survey-based estimator in predicting FGT poverty measures. We focus on the poverty headcount ratio and the poverty gap with three generic poverty lines that render 25%, 50%, and 75% of the population poor. The simulation setting is based on Tarozzi and Deaton (2009). In particular, the target population in the census is a village with $N = 15,000$ households, divided into 150 clusters $k_c \in \{1, \ldots, 150\}$, each of size 100. In each simulation run, an artificial household survey is drawn from the census by selecting randomly ten households from 100 randomly selected clusters. First, both data sets are generated by the following process with homoscedastic errors:

$$
\begin{aligned}
&y_{ch} = \beta_0 + \beta_1 x_{ch} + \eta_c + e_{ch} = 20 + x_{ch} + \eta_c + e_{ch} \\
&x_{ch} = 5 + 0.01 k_c + w_{ch} - t_{ch}, \qquad w_{ch} \sim N(0,1), \qquad t_{ch} \sim U(0,1), \\
&\eta_c \sim N(0, 0.01), \qquad e_{ch} \sim N(0,1).
\end{aligned}
$$

Note that 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 exactly models the data generating process. A linear regression based on the target population yields an R^2 of 0.55 while the new method with an R^2 of 0.08 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 which 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

$$
x_{ch} = 5 + 0.01 k_c + w_{ch}, \quad w_{ch} \sim N(0,1),
$$

where the sampled 100 clusters in the survey have the same values for k_c as they have in the old 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.

Note that in both settings, estimators purely based on the survey have desirable properties as the surveys are representative of the respective village population at the time of data collection. In real-word situations, however, a survey is not necessarily representative at the village-level.

Table 1: *Monte Carlo simulation setting 1 - simultaneous census and survey collection, some variation in the explanatory variable between clusters*

		New estimator			ELL estimator			Survey est.
	True value	Bias	RMSE	Coverage	Bias	RMSE	Coverage	RMSE
$W_0(.25)$	0.2500	0.0025	0.0121	0.9800	0.0017	0.0081	0.9833	0.0137
$W_0(.50)$	0.5000	0.0062	0.0146	0.9767	0.0058	0.0102	0.9633	0.0159
$W_0(.75)$	0.7500	0.0028	0.0113	0.9800	0.0036	0.0084	0.9600	0.0144
$W_1(.25)$	0.0094	0.0000	0.0007	0.9500	-0.0000	0.0005	0.9700	0.0007
$W_1(.50)$	0.0240	0.0002	0.0012	0.9833	0.0001	0.0008	0.9800	0.0012
$W_1(.75)$	0.0473	0.0003	0.0015	0.9800	0.0002	0.0010	0.9900	0.0015

The RMSEs is the root of the mean squared deviations of the estimates from the true value over 300 replications. Coverage rates are calculated for 95% bootstrap percentile intervals.

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, i.e., both cluster effects and household errors are independently sampled with replacement from their sample analogs from the first-stage regression. See Section 2.3 for details.

In the first setting, the root mean squared error is, as expected, smallest for the ELL method, followed by our 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 our new 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% and the bias is negligible.

In the second and 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 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 the magnitude of poverty. In contrast, the new 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 it is the case in this simulation setting) and census and survey size differ distinctly. The latter is often true in practice.[8]

[8]Note that under the stated conditions, our estimator performs better only in predicting the true value on average. In a single sample, the pure survey mean is superior to our 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. But our new method is still prone to idiosyncratic and (small) simulation error, even under correct model specification.

Table 2: *Monte Carlo simulation setting 2 - dated census and recent survey, some variation in the explanatory variable between clusters, explanatory variable changes over time*

		New estimator			ELL estimator			Survey est.
	True value	Bias	RMSE	Coverage	Bias	RMSE	Coverage	RMSE
$W_0(.25)$	0.2500	-0.0035	0.0128	0.9533	0.1186	0.1190	0.0000	0.0155
$W_0(.50)$	0.5000	0.0040	0.0144	0.9833	0.1374	0.1377	0.0000	0.0166
$W_0(.75)$	0.7500	-0.0011	0.0112	0.9867	0.0925	0.0927	0.0000	0.0154
$W_1(.25)$	0.0089	-0.0001	0.0007	0.9367	0.0065	0.0066	0.0000	0.0008
$W_1(.50)$	0.0234	-0.0002	0.0011	0.9767	0.0115	0.0116	0.0000	0.0012
$W_1(.75)$	0.0456	-0.0001	0.0014	0.9833	0.0154	0.0155	0.0000	0.0016

The RMSEs is the root of the mean squared deviations of the estimates from the true value over 300 replications. Coverage rates are calculated for 95% bootstrap percentile intervals.

5 Application to census data from Brazil

In order to test the proposed method in a real-world example, 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 about monthly income at the level of the individual. In addition, the data sets provide 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, and 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 the true ones based on the entire recent census.

The data sets are extracts from the respective censuses. Roughly ten million individuals are included in each data set, corresponding to 6 and 5 percent of the population in 2000 and 2010, respectively. The country is divided into 25 states and 1,980 municipalities. These municipalities constitute the smallest geographical unit which can be matched between 2000 and 2010. Accordingly, we consider them as clusters in the terminology used in the previous sections. Thus, we use averages over 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.[9] The poverty line is set to \$5.5 in 2011 PPP per person and day.[10] For the sake of illustration, we focus on one single Brazilian state, Minas Gerais. In comparison to other states, it features a large number of municipalities (282) which we can match over the two censuses. The data sets comprise 303,134 and 359,051 observed households in 2000 and 2010, respectively, with full information on the used variables. Maintaining the ratio of number of households, we sample randomly about 18,188

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

[10] The World Bank calculates poverty rates at three poverty lines for Brazil, 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`. We chose the highest one since otherwise there are very few households below the other two poverty lines in both years. Our main aim is to illustrate the method's applicability even in settings in which the time span between the data sets is large and relevant changes in the welfare status have occurred over time.

Table 3: *Regression results - new estimator using all households from 2010 census*

Dependent variable: Income	Coefficient estimate	95% confidence interval
Phone	0.448	[0.318; 0.579]
Employment status	-0.518	[-0.668; -0.367]
Urban	0.233	[0.126; 0.340]
Education	0.335	[0.248; 0.422]
Household members	-0.159	[-0.188; -0.130]
Constant	2.655	[2.449; 2.861]
Number of census households	21,543	
Number of municipalities	282	
R^2	0.0950	

households (year 2000) and 21,543 (year 2010) from the respective censuses and treat the resulting data sets as new censuses. The reason for that is not only computational convenience but also the fact that the state of Minas Gerais is the small area of interest and should therefore exhibit a population size similar to common empirical applications in small area estimation. The true headcount ratios in these artificial censuses change substantially over time, from 0.27 percent in 2000 to 0.11 percent in 2010.

As variables with sufficient variation between municipalities and power to explain variation in income we use location (urban or rural), number of household members, availability of a phone as well as employment status and level of schooling completed of the person with the highest educational attainment in the household. When all households from the 2010 census are used, a linear regression with these explanatory variables yields an R^2 of 0.095. The estimates of the regression coefficients can be found in Table 3. We also added squares of the variables, interactions and many other variables to this simple model without obtaining a substantially higher predictive ability measured by the Akaike Information Criterion. 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 0.88.

We draw artificial surveys from the 2010 census by first sampling randomly without replacement 100 municipalities and then sampling without replacement 10 households randomly from each of those municipalities, resulting in an overall survey sample size of 1,000 households. As the number of households differs between municipalities, the estimation at the first stage has to account for these differences by using appropriate weights. Note that this requires knowledge of the number of households in the municipalities at the time of the survey. In practice, when no recent census is available, the number of households at the cluster level can be obtained from a listing exercise which is usually also needed for the sampling scheme for the household survey.

We use a weighted linear regression in the first stage. Means of the explanatory variables over municipalities for the year 2000 are used to explain household per capita income in 2010. 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 to the non-negligible amount of zero income values.[11]

In the second-stage bootstrap procedure, the regression coefficients are sampled from a multivariate normal distribution where the expected values are the first stage estimates and the robust variance-covariance matrix accounts for correlation within the clusters. The error components are generated by a nonparametric bootstrap. In particular, cluster effects are drawn with replacement from the 100 first-stage estimates. The household errors are drawn with replacement from the first-stage residuals belonging to this specific cluster. See also Section 2.3.

For computing an overall state-level poverty measure, it is crucial to know at least approximately the distributions of households over municipalities in the population at the time of the recent survey: The proposed approach imputes poverty measures for the municipalities by using the dated census households. Clearly, a composite measure of those single poverty measures has to account for the number of households in the municipalities at the time of the recent survey.[12]

We compare the performance of our estimator for the headcount ratio[13] in the state of Minas Gerais with the ELL estimator and a simple (weighted) mean based solely on the recent survey. Note that the sample is, in contrast to many real-world applications, representative and rich at the small-area level such that the weighted survey mean is an unbiased poverty estimator by construction. For the ELL first-stage regression, the same explanatory variables are used, yet on the household level and using the 2010 survey data. In a regression based on all households from the 2010 census, this simple model specification already yields an R^2 of 0.33. We conduct 300 Monte Carlo simulations with 200 bootstrap census data sets generated in each replication.

For our estimator, the coverage of the confidence intervals is below the nominal one of 95% (Table 4). The estimator is slightly biased which may be because of unmodeled heterogeneity in the error distribution, e.g. between clusters. In a regression based on all households from the 2010 census, variances and skewness of the residuals differ considerably between clusters (Figure 1). However, we found no clear pattern with respect to the fitted values from a first-stage regression or other explanatory variables. As the number of clusters is relatively small, already one cluster with an extreme behavior of its errors can potentially have a large effect on estimates of welfare measures. In practice, it can be challenging to detect and model such peculiarities in the error distribution. Potential remedies are discussed in Section 6.

Due to the bias in the headcount ratio estimator, a comparison with a (weighted) mean purely based

[11] The proportion of all households in the 2010 census data with an income of zero amounts to 3.16 percent.

[12] In fact, this requirement ensures that changes in the distribution of the explanatory variables are accounted for in our approach. While it is not guaranteed to know the distribution of households at the time of the survey, it is arguably much more realistic than assuming the distribution of the explanatory variables on the household level not to change over time, as done by EEL, for instance.

[13] We also estimated the poverty gap in the same simulation setting and obtained qualitatively similar results.

Figure 1: *Histograms of household residual variances and skewness in clusters*

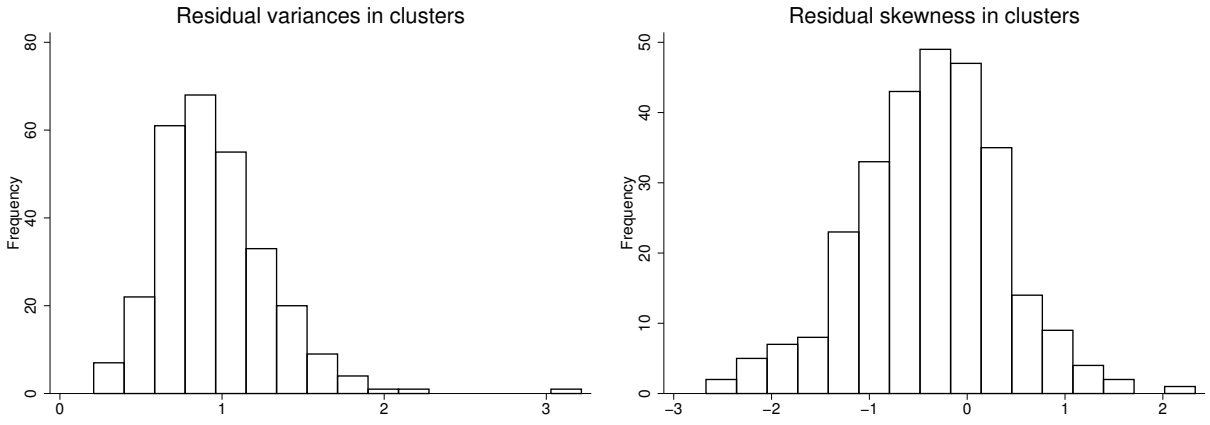

Table 4: *State level headcount ratio at household-level*

		New estimator			ELL estimator			Survey est.	
	True value	Bias	RMSE	Cov.	Bias	RMSE	Cov.	Bias	RMSE
$W_0(5.5)$	0.1076	0.0098	0.0138	0.8900	0.1020	0.1038	0.0000	-0.0015	0.0137

on the survey yields a comparable, even slightly superior performance of the latter in terms of the root mean squared error. Since the distribution of the explanatory variables has changed from 2000 to 2010 (e.g. the share of households owning a phone increased from 67% to 70%), the ELL estimator is severely biased.

So far, the poverty measures have been calculated at the household-level, while one is typically also interested in poverty measures at the individual-level, e.g. the percentage of poor people and not households in a small area. In principle, one could conduct the first-stage regression at the individual level which is equivalent to replicating the household entries in the data sets by the respective household sizes.[14] However, when calculating an overall poverty measure from the simulated income values in the second stage, one then needs to know the number of individuals in each cluster at the time of the recent survey. The required information may be available from a previous listing exercise.

A second option starts with the first-stage regression on the household-level as described above. The smallest unit to match between the census and the survey are the municipalities. In fact, the same value of consumption is predicted on average for all households in the same municipality. For a single bootstrap simulation, they only differ by the simulated household error. Since a relationship between household size and income is assumed on the household level, typically that bigger households are poorer, one cannot randomly assign household sizes to the households. Hence, one possible remedy is to save the household sizes from the survey households and residuals from the first-stage regressions and draw them together in the bootstrap procedure in the second stage.

[14] This is due to the fact that both the household equivalent income and all explanatory variables are the same for all household members.

Table 5: *State level headcount ratio on individual level*

		Our estimator			ELL estimator			Survey est.	
	True value	Bias	RMSE	Cov.	Bias	RMSE	Cov.	Bias	RMSE
$W_0(5.5)$	0.1259	0.0054	0.0126	0.9600	0.1249	0.1270	0.0000	-0.0026	0.0179

Another approach would impute the individual poverty measure based on its relationship with the household poverty estimators. This relationship may be hypothesized on the basis of prior knowledge or estimated from the data set at hand. Though, if the relationship between household sizes and income differs between municipalities, the latter two methods do not yield unbiased state-level poverty estimators in general.

In our application, we follow the second approach, i.e. we run the regression on the household level and sample residuals together with household sizes. The results indicate similar conclusions as the analyses at the household-level (Table 5).

6 Conclusions

In this paper we presented a new method to generate poverty maps.[15] While ours is a valid approach to combine simultaneously collected census and survey data, it also allows analysts to obtain up-to-date poverty maps when only a dated census and a more recent survey are available. In contrast to existing approaches, it has low data requirements and weak assumptions. Simulation studies showed an overall good performance. If the distribution of explanatory variables changes over time, our new estimator is superior to the most frequently used method for contemporaneous census and survey collection.

However, our approach is not immune to issues typically encountered in small area estimation techniques that combine census and survey data. In particular, variable selection and adequate modeling of apparent heteroscedasticity and differences in skewness in the residuals can be challenging. Besides, the key assumption, namely that aggregate household characteristics from the old census relate to consumption the same way in clusters covered by the new survey as in clusters not covered by the new survey, may not hold for the specific welfare estimation exercise at hand. For example, the migration pattern between census and survey collection may vary between clusters and may be correlated with the welfare status which is typically not captured by the model.

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. (2017) 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

[15] Software code in Stata and R for the implementation of our proposed method are available on request from the authors. The recently developed Stata package SAE (Nguyen et al., 2018) can be adapted accordingly.

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.

References

Agostini, C. A., Brown, P. H., and Roman, A. C. (2010). Poverty and inequality among ethnic groups in Chile. *World Development*, 38(7):1036–1046.

Araujo, M. C., Ferreira, F. H., Lanjouw, P., and Özler, B. (2008). Local Inequality and Project Choice: Theory and Evidence From Ecuador. *Journal of Public Economics*, 92(5-6):1022–1046.

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

Bui, T. D. and Nguyen, C. V. (2017). Spatial Poverty Reduction in Vietnam: An Application of Small Area Submission Number. *Economics Bulletin*, 37(3):1785–1796.

Das, S. and Chambers, R. (2017). 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):1137–1161.

Deaton, A. and Kozel, V. (2005). Data and Dogma: The Great Indian Poverty Debate. *World Bank Research Observer*, 20(2):177–199.

Demombynes, G. and Özler, B. (2005). Crime and Local Inequality in South Africa. *Journal of Development Economics*, 76(2):265–292.

Elbers, C., Fujii, T., Lanjouw, P., Özler, B., and Yin, W. (2007). Poverty Alleviation Through Geographic Targeting: How Much Does Disaggregation Help? *Journal of Development Economics*, 83(1):198–213.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, 71(1):355–364.

Elbers, C. and van der Weide, R. (2014). 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*.

Emwanu, T., Hoogeveen, J. G., and Okiira Okwi, P. (2006). Updating Poverty Maps with Panel Data. *World Development*, 34(12):2076–2088.

Foster, J., Greer, J., and Thorbecke, E. (1984). A Class of Decomposable Poverty Measures. *Econometrica:*, 52(3):761–766.

Gibson, J. (2018). Forest Loss and Economic Inequality in the Solomon Islands: Using Small-Area Estimation to Link Environmental Change to Welfare Outcomes. *Ecological Economics*, 148:66–76.

Guadarrama, M., Molina, I., and Rao, J. N. K. (2016). A Comparison of Small Area Estimation Methods for Poverty Mapping. *Statistics in Transition New Series*, 1(17):41–66.

Haslett, S. J. (2016). Small Area Estimation Using Both Survey and Census Unit Record Data. In *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons, Ltd.

Haslett, S. J., Isidro, M. C., and Jones, G. (2010). Comparison of Survey Regression Techniques in the Context of Small Area Estimation of Poverty. *Survey Methodology*, 36(2):157–170.

Healy, A. J., Jitsuchon, S., and Vajaragupta, Y. (2003). Spatially Disaggregated Estimates of Poverty and Inequality in Thailand. *Massachusetts Institute of Technology and Thailand Development Research Institute*.

Isidro, M. C. (2010). *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*. PhD thesis, Massey University, New Zealand.

Isidro, M. C., Haslett, S., and Jones, G. (2016). Extended Structure Preserving Estimation (ESPREE) for Updating Small Area Estimates of Poverty. *Annals of Applied Statistics*, 10(1):451–476.

Kijima, Y. and Lanjouw, P. (2003). Poverty in India During the 1990s: A Regional Perspective. *Policy Research Working Paper No. 3141, The World Bank*.

Marhuenda, Y., Molina, I., Morales, D., and Rao, J. N. (2017). 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):1111–1136.

Minnesota Population Center (2017). Integrated Public Use Microdata Series, International: Version 6.5.

Molina, I. and Rao, J. N. (2010). Small Area Estimation of Poverty Indicators. *Canadian Journal of Statistics*, 38(3):369–385.

Nguyen, M. C., Corral, P., Azevedo, J. P., and Zhao, Q. (2018). sae: A stata package for unit level small area estimation. *Poverty and Equity Global Practice Working Paper Series No. 177, The World Bank*.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed Effects Models in S and S-PLUS*. Springer.

The National Statistical Coordination Board of the Philippines (2009). 2003 City and Municipal Level Poverty Estimates. Available from `https://psa.gov.ph/sites/default/files/2003%20SAE%20of%20poverty%20%28Full%20Report%29_0.pdf`.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Applied Statistics*, 54:507–554.

Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N. (2017). Data-driven Transformations in Small Area Estimation. *Discussion Paper 30/2017, School of Business and Economics, Freie Universität Berlin*.

Tarozzi, A. and Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *Review of Economics and Statistics*, 91(4):773–792.

The Bangladesh Bureau of Statistics, The World Bank and The United Nations World Food Programme (2010). Updating Poverty Maps: Bangladesh Poverty Maps for 2005. Available from `http://www.wfp.org/sites/default/files/Poverty%20Map%202005%20Technical%20Report.pdf`.