Small area estimation of poverty in four West African countries by integrating survey and geospatial data.

Ifeanyi Edochie[1]
David Newhouse[2]
Nikos Tzavidis[3]
Timo Schmid[4]
Elizabeth Foster[1]
Angela Luna Hernandez[3]
Aissatou Ouedraogo[1]
Aly Sanoh[1]
Aboudrahyme Savadogo[1]

Abstract

The paper presents methodology to generate experimental small area estimates (SAE) of poverty in four West African countries: Chad, Guinea, Mali, and Niger. Due to the absence of recent census data in the four countries, household level survey data are integrated with grid-level geospatial data, which are used as covariates in model-based estimation. Leveraging geospatial data enables reporting of poverty estimates more frequently at disaggregated administrative levels and makes estimation feasible in areas for which survey data are not available. The paper leverages the availability of a recent census in Burkina Faso for evaluation purposes. Estimates obtained with the same survey instruments and candidate geospatial covariates as the other four countries are compared against estimates obtained using recent census data and an empirical best predictor under a unit level model. For Burkina Faso, estimates obtained using geospatial data are highly correlated with the census-based ones in sampled areas but moderately correlated in non-sampled areas. The results demonstrate that in the absence of recent census data, small area estimation with publicly available geospatial covariates is feasible, can lead to large efficiency improvements compared to direct estimation and improve the timeliness of small area estimates.

[1] Poverty and Equity Global Practice, World Bank Group

[2] Development Economics Data Group, World Bank Group

[3]Department of Social Statistics and Demography and Southampton Statistical Sciences Research Institute, University of Southampton

[4]Institute of Statistics, Otto-Friedrich-University Bamberg

1. Introduction

This paper presents the methodology to generate experimental small area estimates (SAE) of poverty in four West African countries: Chad, Guinea, Mali, and Niger, as well as an evaluation exercise using data from another country in the region, Burkina Faso. SAE is a statistical method used to improve survey estimates by integrating survey data with geographically comprehensive auxiliary data (covariates) typically derived from census, administrative, remote sensing, or mobile phone data. Data integration is achieved with the use of statistical models to produce estimates at disaggregated geographic levels that are more accurate and precise than estimates that rely only on direct use of the survey data. More disaggregated estimates are key for a better understanding of how to target interventions for the poorest areas as well as for monitoring the impact of such interventions.

Table 1 illustrates the issues with obtaining poverty estimates at disaggregated geographic levels solely from survey data in the countries under study in this paper. The coefficient of variation estimated (cve) is a common measure used to judge the statistical precision of an estimate[3]. Countries often adopt a maximum threshold for the mean or median cves of the estimates that can be reported, which in practice usually ranges from 0.15 to 0.3. For the countries of focus in this paper, the most recent survey estimates of poverty can be obtained from the 2018 round of the Enquête Harmonisée sur le Conditions de Vie des Ménages (EHCVM)[4] which is available at the regional level for each country. The median cve of regional direct estimates produced by the Horvitz-Thompson estimator ranges from 0.07 to 0.12, which is typically within the acceptable range for publication. However, when we examine direct estimates of the poverty rates for the set of target administrative areas, which is one or two levels below the region, the estimates are too imprecise to publish.[5] The target geography in Chad is the latest unofficial definition of departments provided to use by the National Institute of Statistics, while in Guinea it is the Sous-prefectures, and in Mali and Niger it is Communes. At these levels, the median cve of the direct survey estimates reported in Table 1 exceeds the 0.3 threshold for each country except for Chad, where it is 0.27. In addition, not all the target areas are covered by the surveys, making direct estimation impossible for these areas. While this is the case in all countries, unsampled areas are particularly prevalent in Mali, where less than 40% of the target areas are in the sample.

---

[3] The use of the cve requires care especially when reported in relation to estimates of head count ratios. This is because although in theory -for example under simple random sampling- an estimate of 0.1 or 0.9 should have the same variance, the value of the cve is affected by the point estimate leading to different relative precisions. The use of cves in Table 1 is only for illustrating that direct estimates at the required level of geography are not reliable and not for comparing the precision between direct estimates.

[4] This survey is the main output from the Harmonized Surveys on Household Living Conditions Program of the World Bank and the West Africa Economic and Monetary Union (WAEMU) Commission, which resulted in 10 countries (8 WAEMU members plus Guinea and Chad) collecting household data and constructing household welfare using methodologies that were highly harmonized across all the countries and updated in line with international best practice.

[5] These estimates were obtained using an approximation to the variance of the Horvitz-Thompson estimator implemented in the R SAE package (Molina and Marhuenda, 2015), which assumes that the second order inclusion probabilities are the product of first order inclusion probabilities.

Table 1: Statistics of poverty estimates in the focus countries.

Country	**Burkina Faso**	**Chad**	**Guinea**	**Mali**	**Niger**
Year of most recent census	2018	2009	2014	2009	2012
Regions					
Number of regions in sample	13	22	8	9	8
Number of regions in census	13	23	8	9	8
Median cve of sample estimates of headcount poverty rates for regions	0.114	0.123	0.069	0.085	0.079
Target area					
Name of target area	Commune	Department	Subprefecture	Commune	Commune
Number of target areas (population)	351	112	343	704	266
Number of target areas (sample)	234	99	251	244	228
Median cve of sample estimates of headcount poverty rates for target areas	0.435	0.271	0.370	0.415	0.425

Notes: Sample estimates are obtained from the 2018 round of the EHCVM in each country. Median cve of the sample estimates refers to the median estimated coefficient of variation across target areas. Departments in Chad are defined using an unofficial shapefile provided by

the National Institute of Statistics, Economic, and Demographic Studies (INSEED). Survey estimates are based on the subsample of households with valid GPS coordinates.

Typically, small area estimation applications combine survey data with covariates from census (or other population) data. However, except for Burkina Faso, the last time a census was conducted in these countries was between 2009 and 2014. Using out-of-date census data to update small area estimates can lead to biased estimates for example, if the distribution of the census covariates used for prediction has changed over time. This is an issue that is often not discussed in applied poverty mapping work. Literature on approaches to update poverty estimates in the intercensal period includes Isidro et al. (2016), Koebe et al. (2022) and Arias-Salazar (2023). In this paper we rely on using contemporaneous geospatial covariates, as first illustrated by Battese et al. (1988) (see also Nguyen, 2012), to produce small area poverty estimates in countries that lack recent census data.

Advances in processing of geospatial data and the richness of geospatial data sources make their use as auxiliary information in small area models appealing. Newhouse et al. (2023) summarizes recent literature on the use of geospatial data for small area estimation of wealth and poverty. Jean et al. (2016), Yeh et al. (2020), and Chi et al. (2022) show that satellite data are predictive of wealth indices. The present paper utilizes a method commonly used in small area estimation based on the empirical best predictor (EBP) under a nested error regression model (also referred to as mixed model) (Molina and Rao, 2010). When applied to predicting headcount poverty rates using geospatial covariates, this method has yielded predictions that are highly correlated with up-to-date census-based estimates in Mexico, Sri Lanka, and Tanzania (Masaki et al., 2022; Newhouse et al., 2022). The methodology we use in this paper deviates from the official approach endorsed by the World Bank's Poverty Global Practice, as described in Corral et al. (2022), which is based on the EBP under a unit (household) level mixed model, and census micro-data as covariates (referred to as census-EBP). The main difference, besides the use of geospatial (instead of population census) covariates, is that our modelling approach utilizes only grid cell covariates, but the outcome is still modelled at the unit (household) level. This is why sometimes this latter model is referred to as the unit context model.

We explore the use of the unit context model in Chad, Guinea, Mali and Niger that lack recent census data. We further leverage the availability of recent census data for a fifth country in West Africa, Burkina Faso, to conduct an evaluation exercise. The evaluation exercise compares estimates of headcount poverty rates obtained with a unit level model and the empirical best predictor using census covariates, with poverty rates obtained using the empirical best predictor under the unit context model with geospatial covariates.

As noted above, an alternative approach to small area estimation using geospatial covariates is to use an area level model (Fay-Herriot, 1979), case in which both the direct estimates of poverty rates and the geospatial covariates are aggregated at the target area level. Hence, in the evaluation exercise presented in Section 4 we also produce estimates under a Fay-Herriot model as a way of providing additional evidence about the validity of the estimates produced under the unit context model.

Using geospatial data instead of census data in SAE and the unit context model has been criticized in recent literature (for example, Corral et al., 2021). This is due to the possible introduction of omitted variable bias (relative to the unit level model) resulting from the aggregation of the geospatial covariates. Although a detailed discussion of this issue is beyond

the scope of the current paper, being cognizant of the potential impact of using the unit context model on small area estimates is important.

First, the bias that has been reported in the literature is relative to an assumed gold-standard unit (household) level model and the availability of up-to-date household level census micro-data. It is our view that if recent census data are available, the census-EBP method should be preferred. We argue, however, that in the absence of recent census data, the use of geospatial covariates may constitute a valid alternative for providing up-to-date small area estimates until data from the next census becomes available. Second, we have observed that the extent of bias in the unit context model depends on the method used to account for sample weights. In this paper, weights are incorporated following Guadarrama et al. (2018). This weighting procedure was implemented in a way that adjusts the estimates of the regression coefficients and random effects to account for sample weights but does not account for weights when estimating the variance components.[6] As shown below, this can cause significant differences in small area estimates which are larger for models with lower predictive power, which is typically the case with unit context models. In Section 4 we explore both weighted and unweighted versions of the unit context model to assess how this impacts the estimates. Third, noting that aggregation is unavoidable due to the way geospatial data are processed, it is worth mentioning that the geographic level at which geospatial covariates are processed and linked to survey data (grid size) impacts the estimation. Because of this and because geospatial covariates can only act as proxies for the kind of variables typically used to model income (or consumption), it is reasonable to assume that the unit context model may show lower levels of predictive power and higher uncertainty than the unit (household) level model. However, because the estimators of interest are aggregations of individual level predictions, it is not obvious that the lower predictive power and higher uncertainty will substantially reduce the quality of the small area estimates obtained by using the unit context model. Finally, as is the case with any model-based method, model building, variable selection and residual diagnostics are critical. The data analyst can try to mitigate the impact of aggregation by processing the geospatial data at as a fine spatial level as possible to maximize the effective sample size. However, this may increase the risk of observing outliers in the geospatial data. The use of transformations may help make the data more consistent with the assumptions that the functional form is linear, and the error terms are distributed normally. As always, the use of model-diagnostics is crucial.

In addition, Corral et al. (2021) report concerns with the estimated measures of uncertainty under a unit context model. From our perspective, if the model assumptions are satisfied, a parametric bootstrap MSE estimator will provide a valid estimator of the uncertainty under the assumed model. Since the true data generating process is unknown, we cannot know a-priori the extent to which the model assumptions are violated, regardless of the type of model assumed. In Section 4, we present results from Burkina Faso comparing coverage rates derived from the parametric bootstrap under the unit context model, treating census-based estimates as truth, with those from other estimators. For sampled areas, coverage rates under the unit context model for sampled areas are slightly below those from direct estimates and slightly above those obtained from an area level model, indicating that the estimated measures of uncertainty obtained through the parametric bootstrap are reasonable in this case.

In summary, we prefer to avoid making definitive statements about whether the unit context model works well or poorly. We instead posit that in the absence of a recent census, a unit context model with geospatial data may be considered as an alternative to the use of outdated

[6] "For the pseudo-EB estimator, we used the weighted estimator $\hat{\beta}$ given in You and Rao (2002) and the REML estimators of $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$." (Guadarrama et al, 2018, p.8)

census data. The presence of a recent census in Burkina Faso provides a valuable opportunity for evaluating this method. As with every SAE application, the performance of different methods will depend on the country context and the characteristics of the available survey and auxiliary data they are applied to. Evaluations of the estimates therefore remains of paramount importance.

The paper is organized as follows. Section 2 describes the data sources and the process of integrating geospatial and survey data. Section 3 presents the core of the small area methodology, model selection and assessment, small area estimation, mean squared error estimation and measures to assess the small area estimates for all countries of focus in this paper. Section 4 presents an evaluation exercise using recent census and survey data in Burkina Faso. This allows us to compare small area estimates produced with geospatial covariates to small area estimates produced using covariate information from census micro-data. The results of the evaluation exercise add new insights to the body of literature on the use of geospatial data in small area estimation and motivate the use of the unit context model with geospatial data in the four remaining countries that lack up-to-date census data. Section 5 presents experimental point and uncertainty estimates for all countries using the unit context model. The paper concludes with a summary of the main findings and areas for further research.

2. Data sources and geospatial data integration

In this paper, we use geospatial covariates because, as shown in Table 1, the most recent censuses in the four focus countries were conducted in 2014 in Guinea, 2012 in Niger, and in 2009 in Chad and Mali. If more recent census data existed, using these data would be the preferred option. For example, several variables routinely collected in censuses such as household size, education, and sector of employment have been shown to be highly predictive of household welfare. Estimates based on recent census data are expected to be more accurate and precise than estimates based on geospatial data, which is often only available at an aggregated level (see for example Corral et al., 2021).

In this paper, however, we avoid using household level predictors in the model because information for the same predictors from a recent census is not available. Using old census data can be problematic because it is not guaranteed to capture developments since the last census, especially in countries impacted by rapid changes. Interpreting the estimates as if these arise from the census year requires that the distribution of the census predictors, as well as their relationship to poverty, has not changed over time. This is a particular concern in countries such as these under study in this paper, which have among the highest fertility rates in the world and, in addition, have suffered from recent conflict and climate shocks which likely affected the geographic distribution of poverty and the geographic distribution of the population. Alternative sources of administrative data, such as health, land, or other administrative records, can also be useful sources of auxiliary data for small area estimation. However, these were not possible to obtain, and would not necessarily be commonly available for all four countries. We therefore decided to use publicly available, up-to-date geospatial data as covariates in small area models. The full list of candidate geospatial covariates, as well as a brief description of each of them are included in Table A1 in Appendix A.

To estimate the model, we use survey data from the 2018 EHCVM surveys in the focus countries. The process of integrating the geospatial covariates with the survey data in each

country is as follows. First, we process the covariates on a gridded shapefile with square grid cells of size 1 sq km covering the totality of the country. Then, each household in the survey is matched to a grid cell using the centroid of the Enumeration Area in which the household is located. For each country it was observed that in the 2018 EHCVM surveys, geocoordinates were not available for a small share of households (representing less than 7% of all surveyed households in all cases). We dropped these households from the data. A detailed description of the differences between the full sample and the portion with available geocoordinates that was used in the analysis is presented in Table A2 in Appendix A.

Figures 1a and 1b illustrate the use of grid cells and creation of geospatial zonal statistics. Figure 1a shows the grided cells in Conakry, Guinea. Figure 1b shows the value of the average radiance of nighttime lights across grid cells in the same area. The lighter grid cells have higher values of nightlights, while the darker cells have lower values. For each grid cell, we calculated the average feature value from the raster data. In addition to these grid cell-level indicators, we also calculate mean values of the indicator at the target area level to include as predictors in small area models. Including these contextual variables at the target area level as additional covariates helps improve the predictive performance of the model.

Figure 1a: Grid in Conakry, Guinea.

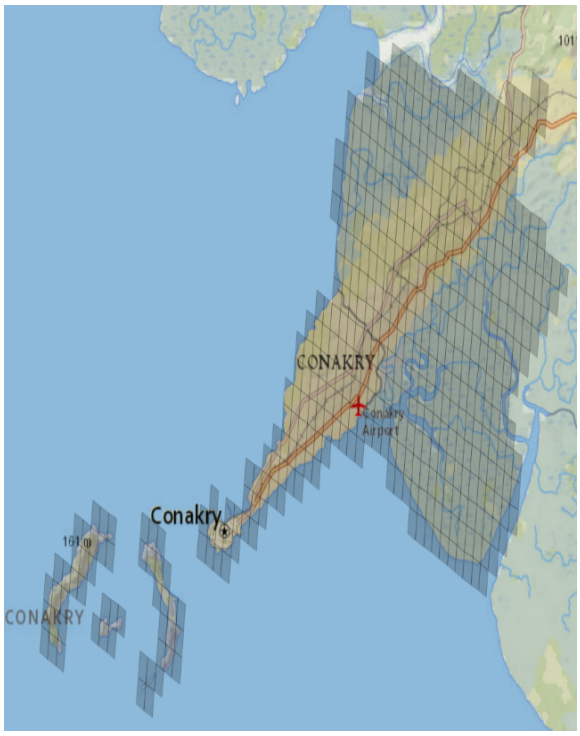
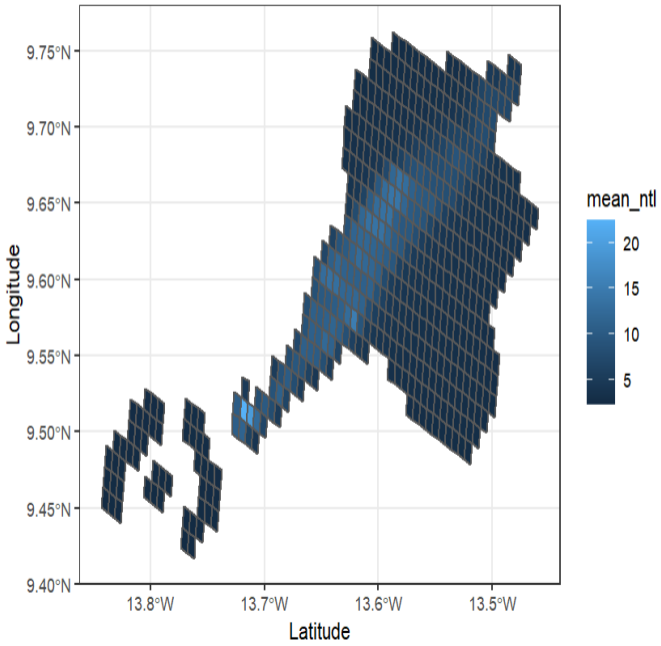


Figure 1b: Average radiance of nighttime lights in Conakry, Guinea



Source: Calculations based on data from Visible Infrared Imaging Radiometer Suite (VIIRS)

3. Small area estimation methodology

In this section we present a summary of the small area methodologies we use to estimate headcount poverty rates at the level of the target area in the five countries of interest. We use a version of the Empirical Best Predictor (EBP) (Battese et al., 1988; Jiang and Lahiri, 2006; Molina and Rao, 2010; Tzavidis et al., 2018) under the unit context nested error regression model with households as the unit of analysis and covariates defined by zonal statistics of geospatial variables at grid cell level (centroid of enumeration areas within target areas). Our methodology is similar to the one used by Masaki et al. (2022), which uses small area estimation to estimate non-monetary poverty indicators in Tanzania and Sri Lanka with geospatial covariates, and Newhouse et al. (2022), which applies similar techniques with geospatial covariates to estimate monetary poverty in Mexico. Van der Weide et al. (2022) also examines the performance of poverty estimates with geospatial covariates in Malawi but using a spatial error model with sub-area level estimates of poverty rates as the outcome, and geospatial zonal statistics as covariates.

As mentioned in previous sections, when the census and survey data are collected from around the same time, using household level census covariates is generally preferred, because the census tends to contain richer auxiliary information than geospatial data. When census data are sufficiently old, however, using cluster level covariate aggregates taken from the old census can generate more accurate estimates than using old census household level covariates (Lange et al., 2021). None of these variations of covariate use, however, reflect any changes in the distribution of the census covariates since the last census. Because, except for Burkina Faso, the census data in the focus countries of this paper are not up to date, we explore the use of more current geospatial data as covariates instead of old census data.

We opt for a household level model of welfare over a grid cell level model of poverty rates because it utilizes more detailed information about the distribution of the welfare variable, and it is easier to interpret. In addition, defining the grid cell level poverty rate as the outcome to be estimated, as in the case of an area level model, requires accounting for the corresponding sampling variability, which may be challenging at such a small level of aggregation. We also prefer a household level model to an area level model because the former allows for the use of auxiliary data at the grid cell level rather than at the target area level, which can improve the accuracy and precision of the estimates, as demonstrated in Masaki et al. (2022) and Newhouse et al. (2022).

We model the household log per capita consumption as a linear function of a subset of geospatial covariates selected through Lasso. The procedure is described in detail in Appendix B. The model equation takes the form:

$$ln\, Y_{ragh} = X_{rag}\beta_1 + X_{ra}\beta_2 + D_r\beta_3 + \nu_a + \epsilon_{ragh}, \qquad (1),$$

where $ln\, Y_{ragh}$ represents the log per capita consumption of household *h*, for which the centroid of their survey enumeration area falls in grid *g* within target area *a* and region *r*. This value of consumption has been spatially deflated using estimated local prices. X_{rag} represents the vector of grid cell geospatial zonal statistics, and X_{ra} represents the vector of unweighted averages of the geospatial variables at the target area level. D_r represents a set of regional dummy variables, ν_a is a random effect specified at the target area level with $\nu_a \sim N(0, \sigma_v^2)$, and ϵ_{ragh} is a household-specific error term with $\epsilon_{ragh} \sim N(0, \sigma_\epsilon^2)$. Survey weights are incorporated into model estimation following Guadarrama et al. (2018), as described in Skarke et al. (2021). A recent paper by Cho et al. (2024) presents optimal predictors for general parameters under an

informative sampling design. Implementing this methodology with the data from the focus countries is a useful area for future research.

The EBP works by repeatedly simulating synthetic populations $ln\ Y_{ragh}$ under model (1) using the expected value of what is unobserved given what is observed in the sample. Under the assumed linear mixed model, this expectation has a closed form. Having fit model (1), the expected log household per capita consumption for each household in the population is computed as follows:

$$ln\ Y_{ragh}^{(l)} = X_{rag}\hat{\beta}_1 + X_{ra}\hat{\beta}_2 + D_r\hat{\beta}_3 + \hat{v}_a + v_a^{(l)} + \epsilon_{ragh}^{(l)} \quad ,(l = 1, .., L),$$

where $\hat{v}_a$ is the random effect predicted with the sample data, $v_a^{(l)}$ is generated from $N(0, \hat{\sigma}_v^2(1 - \hat{\gamma}_a))$, $\epsilon_{ragh}^{(l)}$ is drawn from $N(0, \hat{\sigma}_\varepsilon^2)$ and $\hat{\gamma}_a$is the area-specific shrinkage factor that depends on the estimated variance components and the area sample sizes (Molina and Rao, 2010). For each simulated synthetic population, the target area-specific parameter, the headcount poverty rate, is computed using the simulated values of the welfare variable (per capita consumption) and the official national poverty lines for each country. This procedure is repeated L=100 times and the final estimated poverty rates for each area correspond to the average across the 100 simulations.

In this paper we implement a version of the EBP that calculates the expected value of headcount poverty given the estimated model parameters, for each population unit (which in this case is a grid). Under the assumed linear mixed model, this expectation has a closed form. Having fit model (1), the expected poverty rate for each grid is computed as follows:

$$\hat{P}_{rag} = \Phi\left(\frac{log(Z) - X_{rag}\hat{\beta}_1 - X_{ra}\hat{\beta}_2 - D_r\hat{\beta}_3 - \hat{v}_a}{\hat{\sigma}_v^2(1 - \hat{\gamma}_a) + \hat{\sigma}_\varepsilon^2}\right),$$

where Φ is the standard normal cumulative distribution function, Z is the poverty line, and $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\sigma}_v^2$, $\hat{\sigma}_\varepsilon^2$, and $\hat{v}_a$ are estimated model parameters, and $\hat{\gamma}_a$ is the area-specific shrinkage factor that depends on the estimated variance components and the effective area sample size. The estimated target-area headcount poverty rate, $\hat{P}_{ra}$, is computed by taking a weighted average of the grid-level poverty estimates in each area, with the grid population estimates from WorldPop playing the role of aggregation weights. Estimation is implemented using a modified version of the povmap package (Edochie et al., 2024) in R to implement the estimation.[7] The Povmap package is a modified version of the EMDI package (Kreutzmann et al., 2019) that allows for aggregation weights when aggregating across population units (grids in this case). The two versions of implementing the EBP lead to the same estimates for a large number of Monte-Carlo iterations l . To verify this, we compare poverty headcount and MSE estimates obtained using the traditional method with $L = 100$ Monte-Carlo replications with those obtained by calculating the expected value approach for one focus country, and report the results in Appendix D. Estimates of the mean squared error (MSE) of the small area estimates are calculated using parametric bootstrap under model (1) (Gonzalez-Manteiga et al., 2007) as implemented in the Povmap package. MSE estimation adjusts for the fact that the population data we use contain only one observation per grid, while the actual population contains multiple households per grid. An empirical best predictor under the two-fold version of the nested error regression model is also available (Marhuenda et al, 2017). The two-fold version of the EBP was used to produce official

[7] The code is available on the development branch of the package at: https://github.com/SSA-Statistical-Team-Projects/povmap

small area estimates of poverty rates in Burkina Faso using the latest census in the country. These estimates are used as part of the sensitivity analysis in Section 4.

An alternative approach to small area estimation with geospatial covariates is modelling directly the poverty rates using an area level model (Fay and Herriot, 1979). In this case, both the direct estimates of poverty rates, denoted by $\hat{p}_a^{dir}$, and the geospatial covariates, denoted by X_{ra}, are aggregated at the target area level. The FH model consists of two stages: the sampling model and the linking model. The combination of both stages results in an area level linear mixed model denoted by

$$\hat{p}_a^{dir} = X_{ra}\beta_1 + D_r\beta_2 + \nu_a + \epsilon_a,$$

where X_{ra} represents the vector of unweighted averages of the geospatial variables selected for the model at the target area level. D_r represents a set of regional dummy variables, ν_a is a random effect specified at the target area level with $\nu_a \sim N(0, \sigma_v^2)$ and ϵ_a is the sampling error with $\epsilon_a \sim N(0, \sigma_{\epsilon_a}^2)$. The sampling variance $\sigma_{\epsilon_a}^2$ is estimated under the sampling design and is assumed to be known. The variance component of the random effect is estimated by maximum likelihood methods (e.g., the adjusted maximum-likelihood of Li and Lahiri, 2010) to guarantee positive variance estimates. The MSE of the estimator under the FH model can be obtained by analytic solutions (e.g. Prasad and Rao, 1990) or by bootstrap techniques (Gonzalez-Manteiga et al., 2007). In this paper the FH model is estimated using the Fayherriot Command in Stata (Halbmeier et al, 2019) with no transformation. The routine works similarly to the way the FH model is estimated in the EMDI and Povmap packages (Harmening et al., 2023) in R. Transformed versions of the FH models, using for example the arcsin transformation, are also available and can be considered when modelling proportions. Transformed FH models can be also estimated by using the EMDI and Povmap R packages.

3.1 Model selection and assessment

The geospatial data listed in Table A1 were used to construct averages of zonal statistics both at the grid cell level and the target area level which are used as covariates in model (1). In addition, we include dummy variables at the region level. For model selection we use Lasso to select a set of predictor variables while avoiding overfitting. Estimation of the Lasso penalty parameter is implemented by minimizing the Bayesian Information Criterion (Zhang et al., 2010). The regional dummies are unpenalized and therefore are guaranteed to be selected in the model. Details are provided in Appendix B.

Broadly, the signs and patterns of the coefficients of the unit context model reflect a positive association between population and building density, and a negative association between welfare and remoteness, as proxied by agricultural production and a high prevalence of grassland and shrubland. Fitting model (1) in the five countries under consideration leads to R^2 values ranging from 0.19 in Chad to 0.32 in Niger. This range is consistent with similar applications in other contexts. In similar household level models with aggregation of geospatial covariates at similar spatial scales, the R^2 was 0.30 in Tanzania and 0.27 in Sri Lanka when predicting per capita consumption, and 0.13 in Mexico when predicting per capita income. Geospatial variables do not vary within grids-cells and therefore can only explain variation in welfare across enumeration areas. However, the R^2 is not necessarily the most accurate measure of the benefit of incorporating auxiliary data, as small area estimates based on models with weaker predictors can also be of acceptable quality. Overall, the R^2 values in the focus countries indicate that the geospatial variables measured at grid cell level (enumeration areas) are moderately predictive of

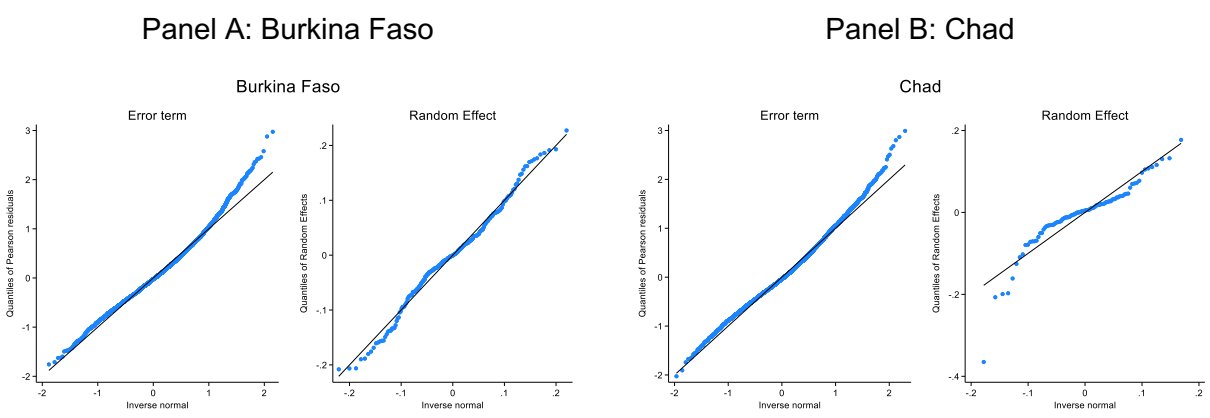
variation in household per capita consumption and can potentially lead to acceptable small area estimates.

Table 2 presents model residual diagnostics under model (1). The error terms appear to be reasonably normal as judged from the skewness and kurtosis, though less so for the unit level error term in Mali. Figure 2 presents quantile-quantile plots for the unit and area estimated model residuals for all five countries. Overall, the results show that the log-transformed model provides a reasonable approximation to the normality of the model error terms. Additional model and residual diagnostics are presented in Appendix C.

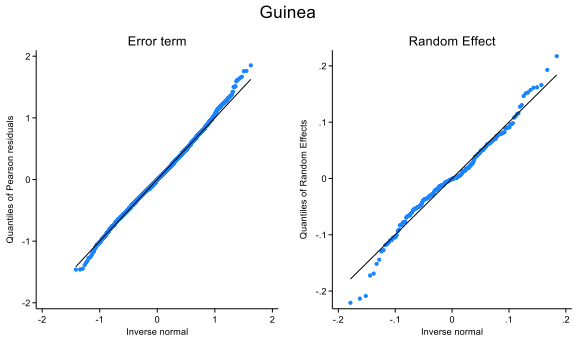
Table 2: Model residual diagnostics.

	Unit level error term		Area effect		Model R^2	
	Skewness	Kurtosis	Skewness	Kurtosis	Marginal	Conditional
Burkina Faso	0.476	3.866	0.194	4.718	0.278	0.392
Chad	0.400	3.477	-0.299	3.408	0.190	0.222
Guinea	0.132	3.310	0.067	4.021	0.272	0.387
Mali	0.559	4.123	-0.093	7.325	0.257	0.330
Niger	0.434	3.644	0.364	4.916	0.317	0.365

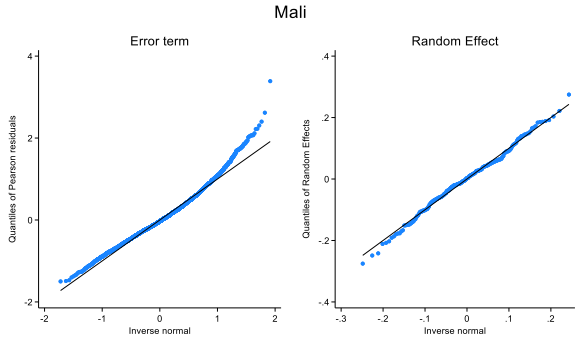
Figure 2: Quantile-quantile plots of unit level error terms and area random effects.



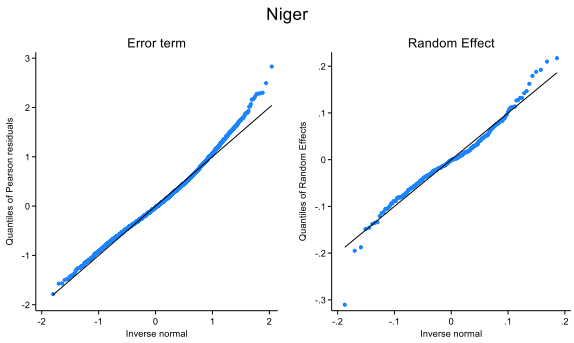
Panel C: Guinea



Panel D: Mali



Panel E: Niger



4. Evaluation exercise: Comparison of geospatial and census-based estimates of headcount poverty in Burkina Faso

Before presenting estimates of headcount poverty rates for the four focus countries that lack recent census data, we conduct a sensitivity analysis with data from Burkina Faso. The availability of a recent census in Burkina Faso creates an opportunity to assess the estimates produced with geospatial covariates and the unit context model against officially adopted EBP census-based estimates as described below.

Burkina Faso's National Institute of Statistics and Demography carried out a census in 2018 which was utilized by the Burkina Faso poverty team of the World Bank to generate small area estimates of poverty for Communes using the EBP census methodology (Molina and Rao, 2010) under a two-fold nested error regression model (Marhuenda et al., 2017). Because the census and the survey data are from a similar period, the small area estimates using census auxiliary information are considered the gold standard. Comparing the census-based estimates with estimates produced using geospatial covariates offers an appropriate testing ground for assessing the extent of discrepancies between census-based and geospatial-based estimates. This framework can also be used to compare the estimates produced by different models.

For the purposes of the evaluation exercise, we treat the census-based EBP estimates as the gold standard. The census-based estimates are compared to: (a) small area estimates under both survey weighted and unweighted versions of model (1) with the outcome defined at household level and the geospatial covariates defined at grid cell level; (b) small area estimates under an area level (Fay-Herriot) model, with geospatial covariates aggregated at the target area level; and (c) small area estimates under a grid cell level model where both the outcome and the geospatial covariates are defined at grid cell level. It is important to note that the survey data was collected from the same harmonized WAEMU survey instrument as the survey data for the other four countries we consider in this paper.

Figure 3 and Table 3 summarize the results of these comparisons. Across all Communes in Burkina Faso, we find a high correlation equal to 0.799 between the estimates under the household level model with geospatial covariates and those derived under the household level model with census covariates. However, there is a large difference in this correlation between in-sample and out-of-sample Communes. For the 234 Communes included in the sample, which comprise 84 percent of the population of Burkina Faso according to WorldPop estimates, the correlation between the survey and census-based estimates is 0.879. In contrast, the correlation for the 117 non-sampled Communes is 0.457. The in-sample correlation is also remarkably similar to findings from other contexts (Masaki et al., 2022; Newhouse et al., 2022; Van der Weide et al., 2022). The correlation for out-of- sample areas meanwhile, is significantly lower than the out-of-sample correlation of 0.7 reported between geospatial and census-based estimates in Mexico (Newhouse et al., 2022). This may be explained by differences in the nature of the geospatial covariates used in Mexico, which could lead to better out of sample predictions, as well as differences in the country context. Perhaps, the lower out-of-sample correlations in this case could be explained by the fact that non-sampled Communes are different from the sampled Communes, as they are more remote, and they are not covered by the survey. The household level, grid cell level, and area level geospatial models all benefit from conditioning on the same household survey data that was used for producing the census-based estimates, making the census and geospatial estimates (household and area level) more consistent with each other in sampled areas. On the other hand, for out-of-sample areas prediction is purely based on grid cell aggregated covariates that may not be as predictive of poverty as household census covariates.

Figure 3: Census-based vs geospatial-based estimates under the unit context model for sampled and non-sampled Communes in Burkina Faso. Red points represent areas (Communes) included in the sample survey, while blue points represent Communes not included in the sample survey.

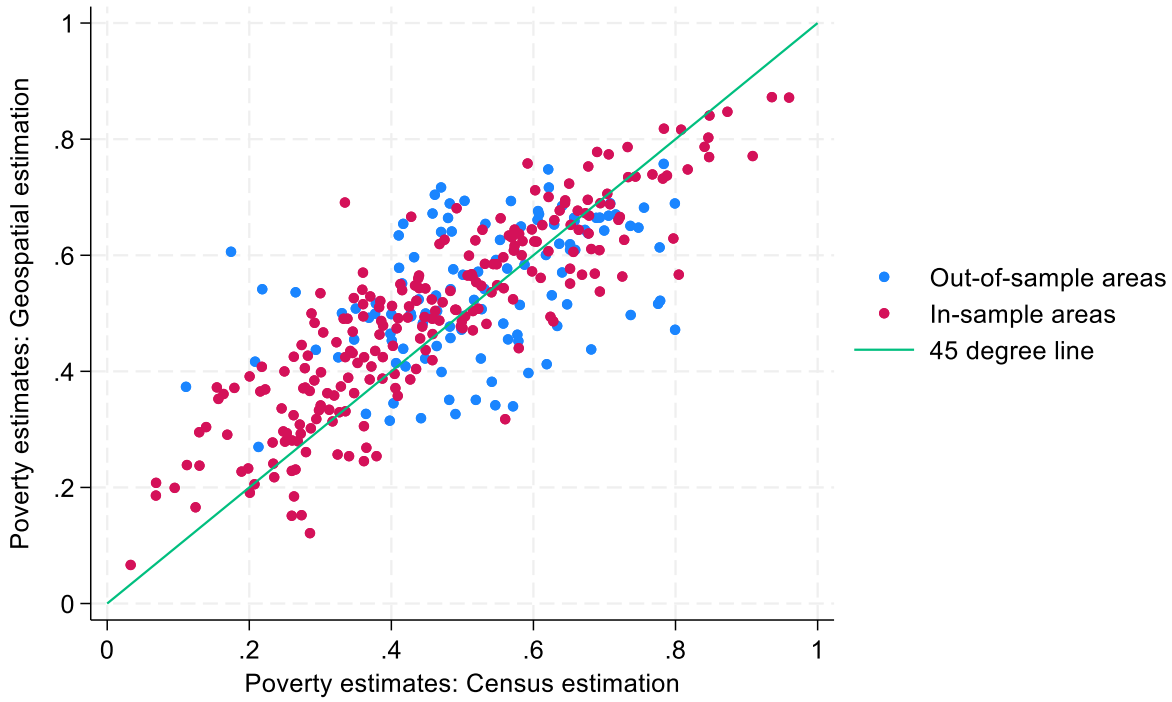


Table 3: Comparison of Commune poverty estimates for different estimation methods in Burkina Faso, by sampled and non-sampled Communes.

	Sampled Communes	Non-sampled Communes	All Communes
Number of Communes	234	117	351
Share of population	83.2%	16.8%	100%
Correlation with census-based estimates			
Household level model with geospatial covariates (with Guadarrama et al. (2018) weights)	0.879	0.457	0.799
Household level model with geospatial covariates (Unweighted)	0.880	0.478	0.807
Grid cell level model with geospatial covariates	0.823	0.529	0.767
Area level model with geospatial covariates	0.754	0.499	0.685
Direct estimates with survey weights	0.837	N/A	N/A
Average estimated MSE across Communes			
Household level model with geospatial covariates (with Guadarrama et al. (2018) weights)	0.007	0.023	0.013
Household level model with geospatial covariates (Unweighted)	0.006	0.023	0.012

Grid cell level model with geospatial covariates	0.015	0.029	0.020
Area level model with geospatial covariates	0.015	0.025	0.018
Direct estimates with survey weights	0.047	N/A	N/A
Coverage rate relative to census-based estimates			
Household level model with geospatial covariates (with Guadarrama et al. (2018) weights)	89.7%	97.4%	92.3%
Household level model with geospatial covariates (Unweighted)	86.3%	95.7%	89.5%
Grid cell level model with geospatial covariates	96.2%	97.4%	96.6%
Area level model with geospatial covariates	86.8%	87.2%	86.9%
Direct estimates with survey weights	91.0%	N/A	N/A
Average estimated headcount poverty rate across Communes			
Census-based estimates	45.7%	52.6%	48.0%
Household level model with geospatial covariates (with Guadarrama et al. (2018) weights)	49.3%	54.1%	50.9%
Household level model with geospatial covariates (Unweighted)	44.0%	47.6%	45.2%
Grid cell level model with geospatial covariates	48.5%	52.7%	49.9%
Area level model with geospatial covariates	39.1%	41.8%	40.0%
Direct estimates with WorldPop weights	48.2%	N/A	N/A
Direct estimates with survey weights	46.8%	N/A	N/A

Looking at the MSE estimates in Table 3, the household level model generates estimates with lower MSEs on average compared to the estimates under the grid cell level and area level models. A further comparison between the estimates produced with geospatial covariates and the census-based ones is to compute coverage rates by treating the assumed gold standard census-based estimates as the truth. The coverage rate is the share of Communes for which a 95% normal confidence interval for headcount poverty, defined as the estimate plus/minus 1.96 times the square root of the estimated MSE, contains the census-based estimate. Overall, the coverage rate for estimates under the household level model with geospatial covariates is 92.3%. Of course, this is not an ideal test because the census-based estimates are themselves estimates, derived from the same sample data as the geospatial based estimates. Nonetheless, the high coverage rate alleviates concerns about the validity of the estimates produced under the unit context model.

Recent research suggests that machine learning methods that allow for more flexible functional forms can improve small area prediction (Krennmair and Schmid, 2022, Merfeld and Newhouse, 2023). Exploring whether the use of machine learning methods improves prediction for out-of-sample areas is an area of current research focus. The performance for out-of-sample prediction will depend on the focus country and how well geospatial data predict poverty. Therefore, out-of-sample estimates under the unit context model in the four countries that lack recent census data should be interpreted with great caution and are likely to change when the next round of census-based estimates becomes available.

5. Assessment of experimental SAE estimates of head count poverty in Burkina Faso, Chad, Guinea, Mali and Niger

Having compared geospatial and census-based small area estimates of head count poverty in Burkina Faso, this section describes the generation of experimental small area estimates of head count poverty for all five countries. Estimates are produced using model (1) with geospatial covariates as described in Table A1. Because the estimates we produce are experimental and not official, the results we present do not identify target areas in the focus countries. Model-based estimates under model (1) with geospatial covariates are compared to direct estimates both at target area and at aggregate, regional levels. In addition, MSE estimates of the model-based estimates are compared to the estimated variances of the direct estimates.

Figure 4 shows the relationship between the EBP estimates under model (1) and the direct estimates at the target area level. In general, model-based estimates are strongly correlated with direct estimates and exhibit less variation than direct estimates, as one would expect due to the impact of shrinkage.

Figure 4: Direct estimates vs. model-based (under the unit context model with weighting following Guadarrama et al., 2018) estimates at target area level.

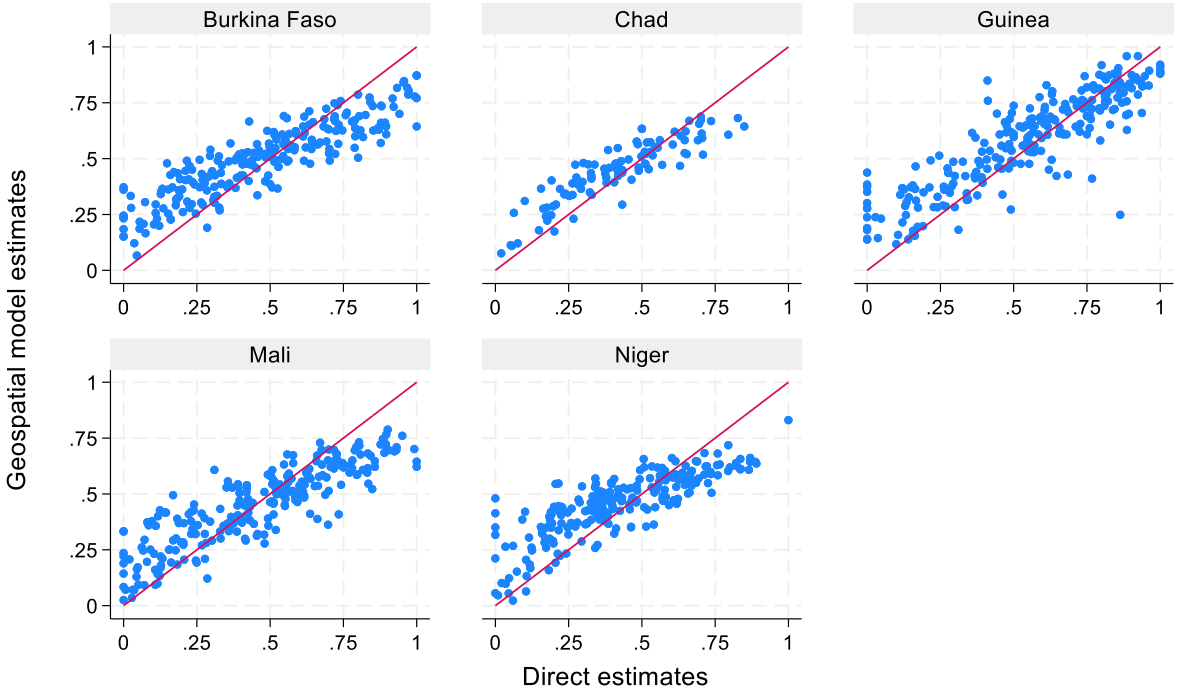


Figure 5 shows the relationship between the EBP estimates under the weighted and unweighted version of model (1). The results show that weighted model-based estimates are systematically higher than the unweighted estimates, while the unweighted estimates are closer to the direct estimates. This may be due to the approach to weighting taken by the Guadarrama et al. (2018) method. In future research it will be interesting to compare the current estimates against estimates obtained by using other weighting methods, including the method that accounts for informative sampling proposed by Cho et al. (2024). For the remaining of this section, we will use the term model-based estimates to refer to those obtained under the unit context model using the weighting proposed in Guadarrama et al. (2018).

Figure 5: Unweighted model-based estimates vs weighted model-based estimates (under the unit context model with weighting following Guadarrama et al., 2018) at target area level

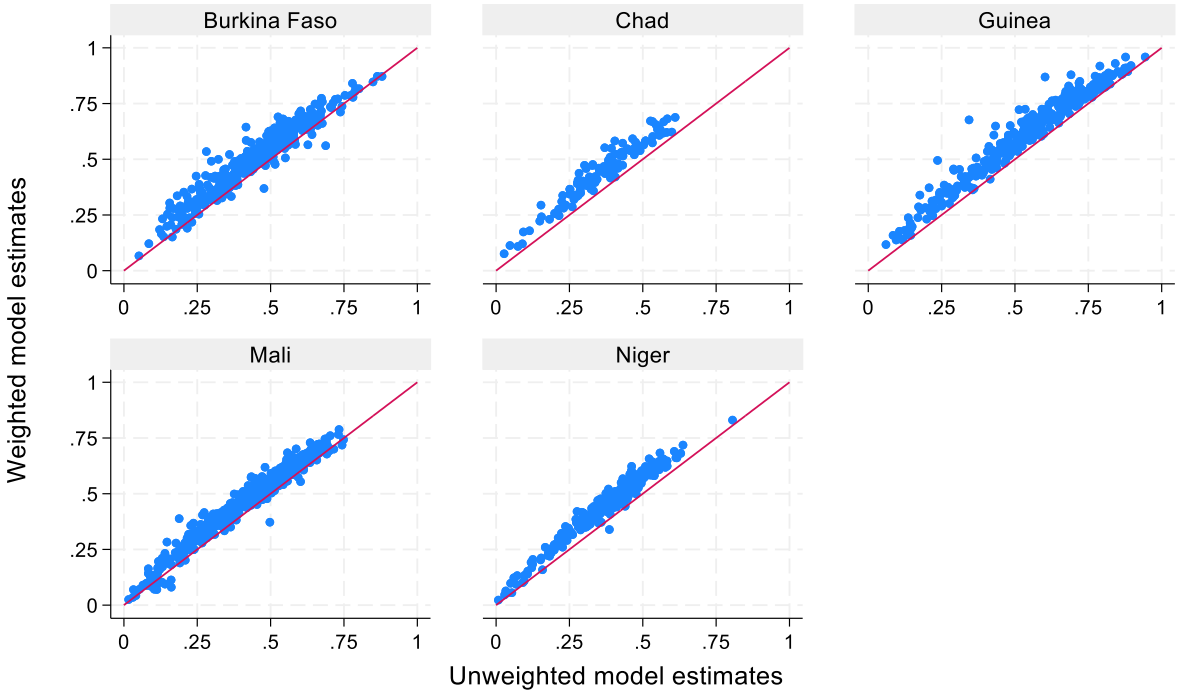


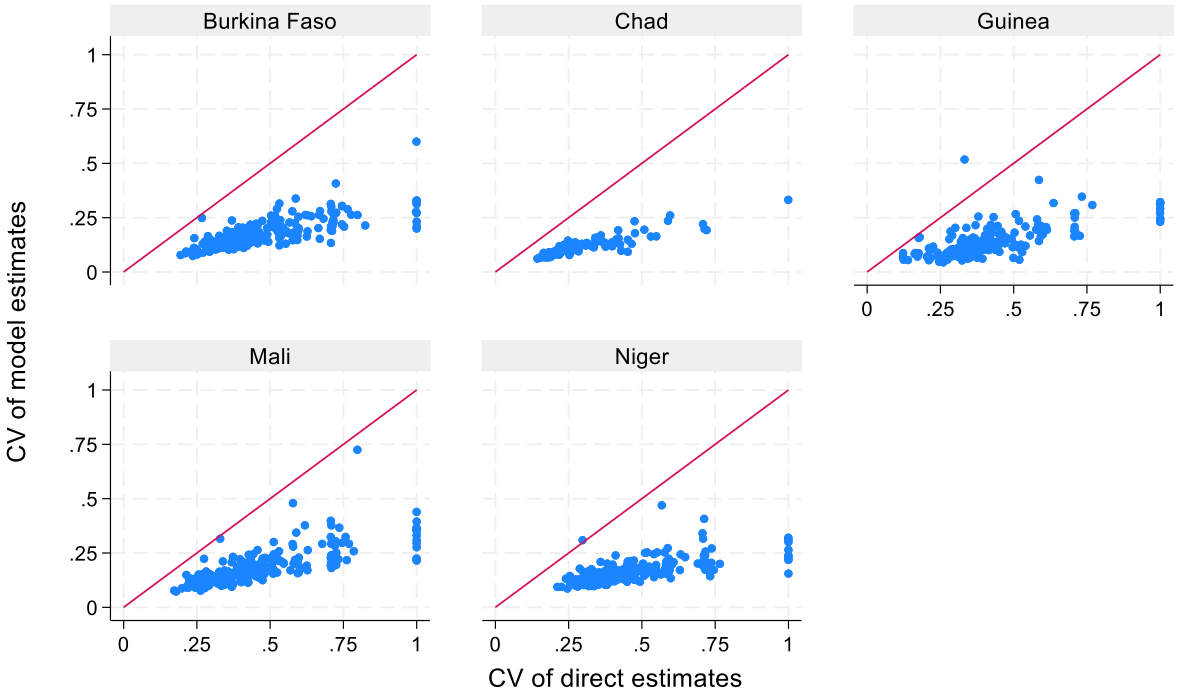
Table 4 presents the median, over target areas, of the coefficient of variation estimated for the model-based versus the direct estimates. Our preferred measure of uncertainty for the direct estimates is based on the Horvitz-Thompson approximation, calculated using the R SAE package, with the sum of the sample weights for each area used to approximate the domain size.

Table 4: Median cve for direct and model-based estimates.

Country	Burkina Faso	Chad	Guinea	Mali	Niger
Direct survey estimates					
Sampled areas	0.435	0.271	0.370	0.415	0.425
Model-based estimates					
Sampled areas	0.167	0.115	0.121	0.164	0.155
Non-sampled areas	0.2084	0.268	0.220	0.229	0.229

Median percentage reduction in cve in sampled areas	62.8%	59.7%	68.3%	60.6%	64.0%

Figure 6: Cve for direct (Horvitz-Thompson approximation) Vs. cve for model-based estimates for sampled areas by country.



The results in Table 4 show a large reduction in the median cve of the model-based estimates relative to direct estimates. Figure 6 shows reductions in the cve for all but a few target areas. The large efficiency gains from the use of model-based estimates are possibly moderately overestimated. In real data evaluations (e.g., Masaki et al., 2022 and Newhouse et al., 2022), coverage rates of confidence intervals produced by using parametric bootstrap MSE estimates are somewhat below the nominal 95%. Nevertheless, even considering this, we expect the model-based estimates to be more efficient than direct estimates.

As a further comparison between model-based and direct estimates, we consider a goodness of fit statistic at the target area level, following Brown et al. (2001). The statistic is based on computing Z scores defined as follows,

$$Z_a = \frac{\left(\hat{p}_a^{ebp} - \hat{p}_a^{dir}\right)}{\left(\sqrt{\widehat{MSE}_{ebp,a} + \widehat{VAR}_{direct,a}}\right)}, \quad (2)$$

Where in (2) $\hat{p}_a^{dir}$, $\hat{p}_a^{ebp}$ are the direct and model-based estimates of headcount poverty rates under model (1) for area a respectively, and the denominator comprises the estimated mean squared error of the EBP estimates and the estimated variance of the direct estimates. The Z scores are useful for assessing the magnitude of the difference between the direct and model-based estimates relative to corresponding uncertainty estimates and whether the differences, taken collectively over all areas, are statistically significant. The Wald test statistic is defined by,

$$W = \sum_a Z_a^2 , \quad (3)$$

where W is distributed as a chi-squared distribution with degrees of freedom equal to the number of areas. A value below the 95 percent threshold implies a p-value above 0.05, indicating that the differences are not statistically significant. Table 5 presents the p-value for each country when using the EBP estimates under the weighted version of the unit context model. For Burkina Faso, Chad and Mali we don't find statistically significant differences between the model-based and direct estimates. However, this is not the case for Guinea and Niger. Meanwhile, Table 6 presents the p-value for each country when using the EBP estimates under the unweighted version of the unit context model. In this case all p-values exceed 0.05, indicating that the differences between the direct and EBP estimates are not statistically significant at the 95 percent level. These results, along with the average poverty estimates reported in Table 3, show the significant impact that weighting can have on model-based estimates and highlight the need for further research on weighting methods.

Table 5: Results from applying the goodness of fit test (3) in the 5 countries (Weighted version of the unit context model).

Country	Test statistic (W)	95% Threshold	Degrees of Freedom	p-value
Burkina Faso	251.7	270.7	234	0.2
Chad	94.4	123.2	99	0.61
Guinea	299.9	289.0	251	0.02
Mali	260.4	281.4	244	0.23
Niger	271.8	264.2	228	0.025

Table 6: Results from applying the goodness of fit test (3) in the 5 countries (Unweighted version of the unit context model).

Country	Test statistic (W)	95% Threshold	Degrees of Freedom	p-value

Burkina Faso	231.1	270.7	234	0.54
Chad	87.3	123.2	99	0.79
Guinea	246.4	289.0	251	0.57
Mali	280.6	281.4	244	0.054
Niger	216.5	264.2	228	0.70

Turning now to comparisons at more aggregate levels, Figures 7, 8 and 9 compare the EBP geospatial estimates with direct estimates at the regional level. Figure 8 shows the same results as Figure 7 but excludes out of sample areas when aggregating the EBP estimates at the regional level. We decided to explore this latter approach considering the lower correlation between unit context model estimates and the unit level model estimates for out of sample areas in Burkina Faso. Both Figures 7 and 8 aggregate the model-based estimates using WorldPop weights, while the direct estimates are aggregated using survey weights. Figure 9 remedies this inconsistency by using survey weights when aggregating the model-based estimates from target areas to regions. Overall, the results show that the model-based estimates are aligned well with direct estimates at the regional level. Some discrepancies are to be expected, however, because model-based estimates are affected by shrinkage.

Figure 7: Small area estimates vs direct estimates at regional level (using WorldPop weights for aggregation)

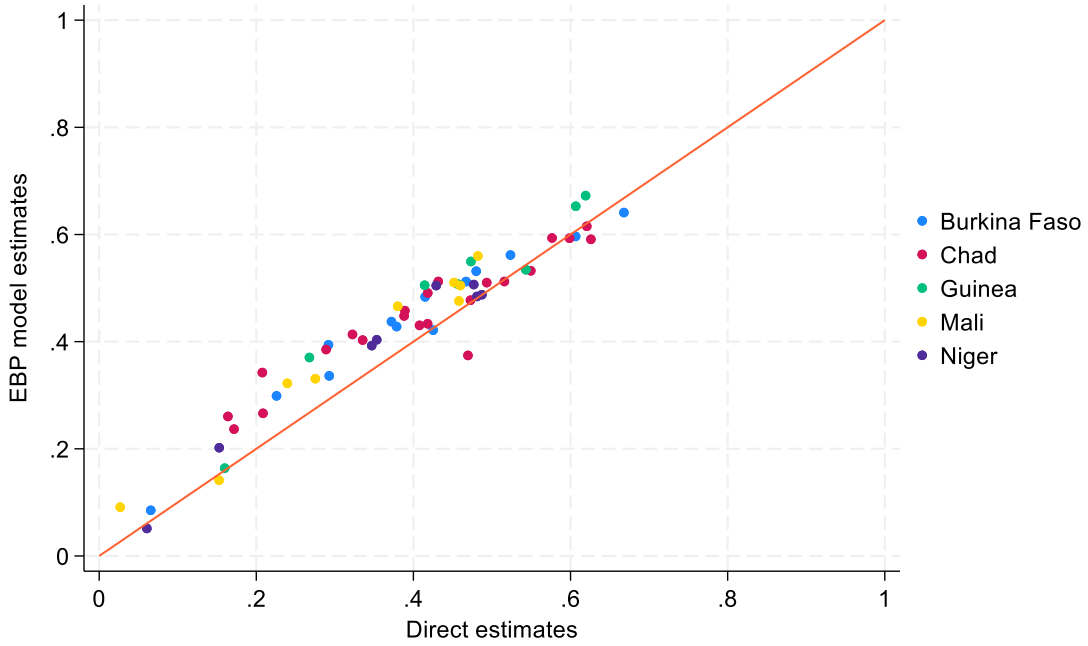


Figure 8: Small area estimates vs direct estimates at regional level (including only in sample target areas and using WorldPop weights for aggregation).

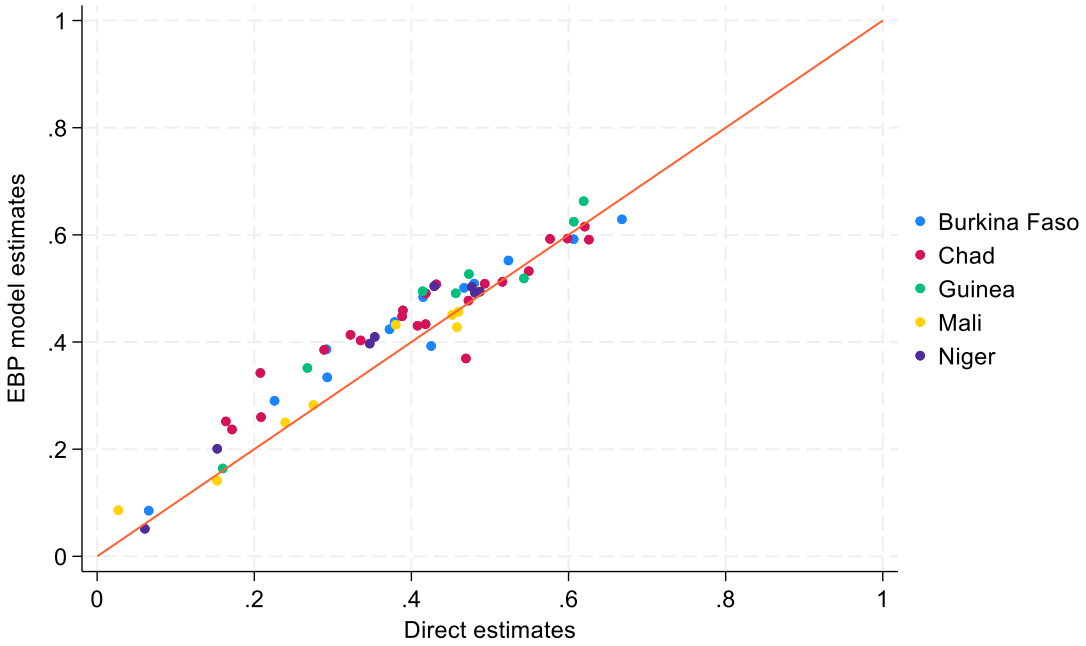
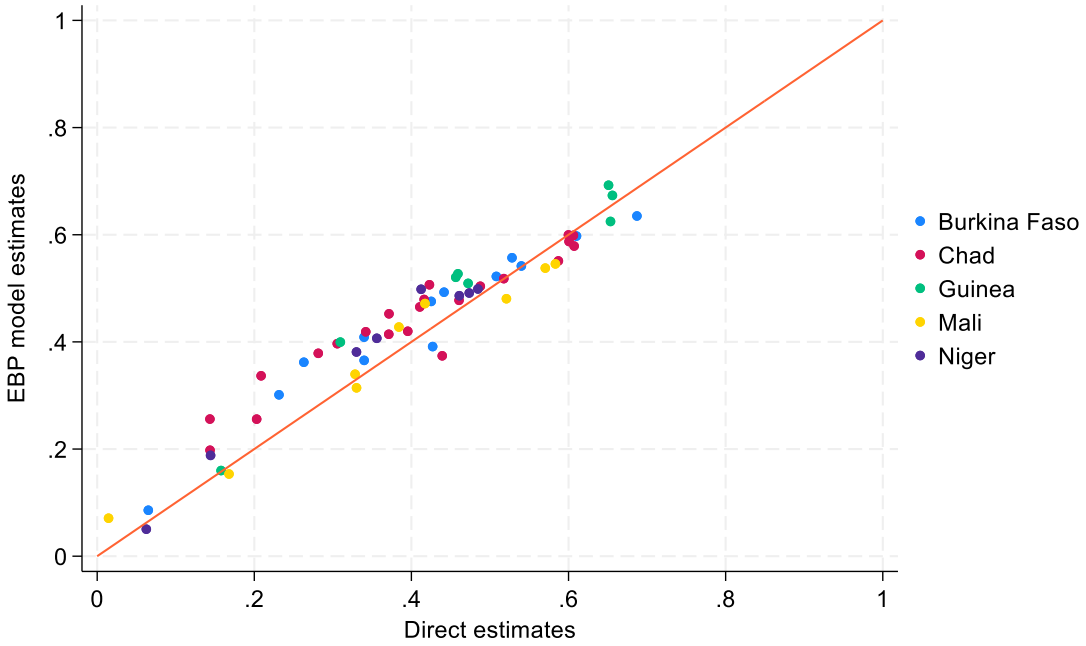


Figure 9: Small area estimates vs direct estimates at regional level (including only in sample target areas and using survey weights for aggregation).



6. Conclusions

This paper describes the methodology used for producing experimental small area estimates of headcount poverty rates in five West African countries where in four of these countries no recent census data are available and zonal statistics from geospatial data sources are available instead. The use of model-based estimation with geospatial covariates offers a pragmatic

approach for producing interim model-based estimates because the alternative of using old census data carries risks especially if the distribution of census variables has changed over the intercensal period.

The presence of a recent census in Burkina Faso provided a valuable opportunity to evaluate the results of different models against 'gold standard' census-based estimates. In sampled areas, the estimates produced by the unit context model track the census-based estimates well and have lower MSEs than direct estimates. Across all areas, the correlation between the geospatial-based estimates and the census-based estimates is high, but this correlation was much higher in sampled areas than non-sampled areas. Models specified at the household level generated estimates that were moderately more accurate than those specified at the grid cell level, because the greater variation in per capita consumption allowed for the automated selection of a richer model. Both sets of estimates had lower MSEs than estimates under a model specified at the area level which we think is due to the use of more granular auxiliary data.

Overall, the estimates for the countries without census data show large improvements in MSE reduction compared to direct estimates. In particular, the median cve in sampled areas is reduced between approximately 59% and 68%. The five countries focus of this paper are neighbors and share many economic and social characteristics. Furthermore, all of them implemented highly harmonized surveys concurrently, and the set of geospatial variables available for model selection is identical. However, one cannot be certain that the results for Burkina Faso generalize to the other four West African countries as there are important differences to take into account. Burkina Faso is facing a significant internally displaced people crisis, affecting about 10% of the population, but hosts far fewer refugees than Niger or Chad. Burkina Faso and Guinea lack the large areas of mainly uninhabited desert that characterize Mali, Niger, and Chad. Nonetheless, the relatively low correlation between the geospatial estimates and the census estimates in non-sampled areas observed in Burkina Faso raises the prospect that the estimates for these areas could significantly change when upcoming censuses are collected and combined with survey data to produce updated poverty maps. Given the scarcity of evidence on out-of-sample prediction accuracy in the literature, we recommend treating the out-of-sample estimates in the remaining four countries with a high degree of caution.

There are several additional avenues for further research to inform these types of data integration efforts. These include additional empirical work to validate both point and uncertainty estimates against estimates using recent census data. Zonal statistics derived from geospatial data can be highly correlated. Initial results from current research indicate that the approach used for geospatial data processing and model building with geospatial data can impact on the quality of the produced estimates. Model estimation also matters, it should be possible to improve upon existing methods for estimating mixed models with sampling weights. In addition, exploring the use of machine learning methods to capture complex relationships could improve estimation especially for out-of-sample areas. Despite room for further improvement, the model-based estimates of the type calculated in this paper can provide interim estimates of the spatial distribution of poverty with acceptable uncertainty measures that cannot be obtained with survey data alone.

References

Arias-Salazar, A. (2023). “Small area estimates of poverty incidence in Costa Rica under a structure preserving estimation (SPREE) Approach, *Journal of Official Statistics*, 39(4), 435-458.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28-36.

Brown G., Chambers R., Heady P., & Heasman D. (2001). Evaluation of small area estimation methods: an application to unemployment estimates from the UK LFS. Proceedings of Statistics Canada Symposium 2001.

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries*. Proceedings of the National Academy of Sciences*, 119(3), e2113658119.

Cho, Y., Guadarrama-Sanz, M., Molina, I., Eideh, A., & Berg, E. (2024). Optimal predictors of general small area parameters under an informative sample design using parametric sample distribution models, *Journal of Survey Statistics and Methodology*, https://doi.org/10.1093/jssam/smae007

Corral, P, Molina, I., Cojacaru A., & Segovia S. (2022). Guidelines for poverty mapping, World Bank.

Corral P, Himelein K, McGee K, & Molina I. (2021). A map of the poor or a poor map? *Mathematics*. 9 (21); 2780. https://doi.org/10.3390/math9212780.

Fay, R. E., & Herriot, R. A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*, 269-277.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, *78*(5), 443-462.

Guadarrama, M., Molina, I., & Rao, J. N. K. (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics & Data Analysis*, *121*, 20-40.

Halbmeier, C., Kreutzmann, A. K., Schmid, T., & Schröder, C. (2019). The fayherriot command for estimating small-area indicators. *The Stata Journal*, 19(3). 626-644.

Harmening, S., Kreutzmann, A. K., Schmidt, S., Salvati, N., & Schmid, T. (2023). A framework for producing small area estimates based on area-level models in R. *The R Journal*, 15(1). 316-341.

Edochie, I., Newhouse, D., Würz, N., & Schmid, T. (2024). Povmap: Extensions to the emdi package, R package version 1.0.1, https://CRAN.R-project.org/package=povmap.

Isidro, M, Haslett S., & Jones G. (2016). Extended structure preserving estimation (ESPREE) for updating small area estimates of poverty. *Annals of Applied Statistics* 10 (1), 451–476.

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

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, *15*(1), 1-96.

Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., & Schmid, T. (2022): Intercensal updating using structure-preserving methods and satellite imagery, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185, 170-196.

Kreutzmann, A. K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, *91*.

Krennmair, P., & Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *71*(5), 1865-1894.

Lange, S., Pape, U. J., & Pütz, P. (2021). Small area estimation of poverty under structural change. *Review of Income and Wealth*, 68, 264-281.

Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101(4), 882-892.

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. N. K. (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.

Molina I, & Marhuenda Y (2015). sae: An R package for small area estimation. *The R Journal*, 7(1), 81–98.

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2022). Small area estimation of non-monetary poverty with geospatial data, *Statistical Journal of the IAOS,* v 37 no. 4

Merfeld, J. D., & Newhouse, D. (2023). Improving estimates of mean welfare and uncertainty in developing countries (Policy Research Working Paper No. 10348). The World Bank.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369-385.

Newhouse, D. (2023) "Small Area Estimation of poverty and wealth using geospatial data: What have we learned So Far?." *Calcutta Statistical Association Bulletin*: 00080683231198591.

Newhouse, D., Ramakrishnan, A., Merfeld, J., Swartz, T., & Lahiri, P. (2022), Small area estimation of monetary poverty in Mexico using geospatial data*, World Bank Policy Research Paper* no.10175.

Nguyen, V. C. (2012). A method to update poverty maps. *Journal of Development Studies*, 48 (12), 1844-1863.

Prasad, N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Skarke, F., Kreutzmann, A. K., & Würz. N. (2021) Extensions to the ebp function in the R package emdi: additional data-driven transformations and empirical best prediction under informative sampling.

Tzavidis, N., Zhang, L. C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(4), 927-979.

Van der Weide, R., Blankenspoor, B., Elbers, C, & Lanjouw, P (2022) How accurate is a poverty map based on remote sensing? An application to Malawi. *World Bank Policy Research Paper* no. 10171.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., ... & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications*, 11(1), 2583.

Zhang, Y., Li, R., & Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*(489), 312-323.

Appendix A: Information on data sources

Table A1: List of candidate geospatial variables.

Variable	Source	Approximate Resolution	Year
Population structure	WorldPop (https://www.worldpop.org)	100 m	2018
Population density	WorldPop	100 m	2018
Temperature	TerraClimate (https://www.climatologylab.org/terraclimate.html)	4 km	2018
Palmer Draught Severity Index (PSDI)	TerraClimate	4 km	2018
Distance to OSM major roads	WorldPop	100 m	2016
Radiance of night-time lights	VIIRS (https://eogdata.mines.edu/products/vnl/)	500 m	2018
Net primary production	FAO Remote Sensing for Water Productivity (WaPOR) 2.1 (https://data.apps.fao.org/wapor/?lang=en)	240 m	2018
Rainfall	Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) (https://www.chc.ucsb.edu/data/chirps)	5.5 km	2018
Elevation	NASA's SRTM Digital Elevation (3 arc seconds spatial resolution) (https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1)	30 m	2018
Cellphone tower count	The OpenCell ID project (https://www.opencellid.org/#zoom=16&lat=37.77889&lon=-122.41942)	1 km	April 2022
Years since change to impervious surface	Tsinghua University via Google Earth Engine (https://developers.google.com/earth-engine/datasets/catalog/Tsinghua_FROM-GLC_GAIA_v10)	30 m	2018
Building count	WorldPop	100 m	2018

Coefficient of variation on buildings	WorldPop	100 m	2018
Land cover classifications	Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 (https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_Landcover_100m_Proba-V-C3_Global)	100 m	2018

Table A2: Differences between full survey sample and subsample with available geospatial coordinates

	Burkina Faso	**Chad**	**Guinea**	**Mali**	**Niger**
Population					
Number of target areas	351	112	343	704	266
Full survey sample					
Number of target areas	N/A [8]	N/A	N/A	N/A	N/A
Number of enumeration areas	585	627	688	551	504
Number of households	7,010	7,497	8,256	6,602	6,024
National poverty rate	41.4	41.9	43.7	41.9	40.8
Survey subsample with geospatial coordinates					

[8] Households without geospatial coordinates don't have the required information to assign them to a target area

Number of target areas	234	99	251	244	228
Number of enumeration areas	555	594	681	531	483
Number of households	6,650	7,124	8,159	6,362	5,777
National poverty rate (subsample)	41.0	42.8	43.9	42.0	40.9

Notes: Only includes completed interviews. National poverty rate taken from World Development Indicators. All other figures taken from staff calculations based on 2018 EHCVM survey data.

Appendix B: Method for model selection

Lasso selects variables using the following optimization problem:

$$\hat{\beta} = argmin\, \beta \in R^p \;\; E_n[(y_i - X_i\beta)^2] + \frac{\lambda}{N}\sum_{j=1}^{p}\left|P_j\beta_j\right| \quad \text{(B1)}$$

Where y_i is log per capita consumption for household i, X_i is a matrix of p normalized candidate predictor variables, R^p represents the set of p-dimensional real numbers, $E_n[\,]$ represents the empirical average in the sample of n households, P_j is a variable specific penalty parameter that takes on the values of zero or one, and λ is a penalty parameter. When lambda is set to 0, LASSO is equivalent to OLS, and as lambda increases, the optimal solution sets an increasing number of coefficients to zero. The penalty parameter P_j is set to zero for the regional dummies, to ensure they are selected, and one for all other variables. The remaining non-zero coefficients are the selected variables. Thus, the value of the penalty factor λ determines how many variables are selected for the model.

We select the value of λ that minimizes the Bayesian Information Criteria, defined as:

$$BIC = ln(2\pi) + ln(E_N[(y_i - X_i\beta)^2]) + 1 + CN \quad \text{(B2)}$$

Where C is the number of non-zero coefficients, X_i is a 1 by C vector of variables and β is a C by 1 vector of coefficients estimated without a penalty. X_i, β, and C are all functions of λ. Larger values of λ are associated with sparser models, which increase the value of the second term but reduce the value of the third term of (B2)

The resulting lambda is plugged into equation (B1) to determine the set of variables with non-zero coefficients. These variables, without further selection, are incorporated into the linear mixed model specified in equation (1).

Appendix C: Small area estimation with geospatial variables

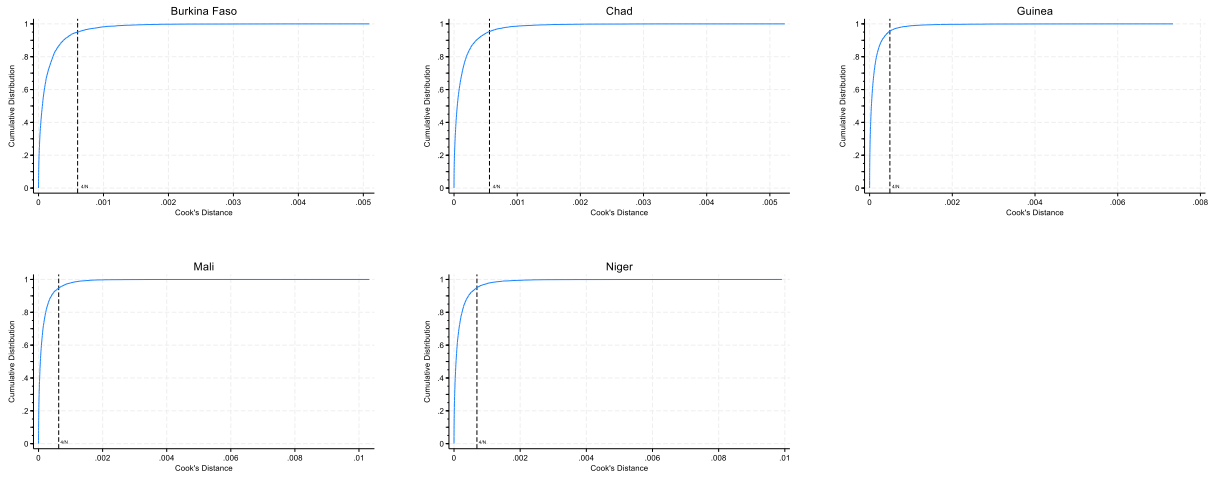
Residual Diagnostics

We look at the impact of influential data points on the model fit by computing Cook's distance. This is estimated by removing the i[th observation] from the data and summarizing the change in the regression model as follows:

$$D_i = \frac{\left(\sum_{j=1}^{n} Y_j - Y_{j(i)}\right)^2}{(p+1)\sigma^2}$$

We estimate D for all 4-unit context models below. A data point (index) is said to be influential by the cook's distance measure when $D_i > \frac{4}{N}$ (where N is the number of observations in the survey), an arbitrary rule of thumb common within the statistical literature.

Figure C1: Cumulative Distribution Functions of Cook's Distance by country



Cross Validation

To ensure that the models are not overfit, we conduct a 10-fold cross validation exercise of the models used in Mali, Guinea, Chad and Niger. The 10-fold cross validation procedure involves randomly selecting 90% of the sample (matching the survey design), training a welfare model on this data sample, predicting the outcome into the remaining 10%, and repeating the process a total of 10 times.

Table C1: In and out of sample R2 from cross-validation exercise.

	Marginal R2	In-sample R2	Out of sample R2
Burkina Faso	0.278	0.321	0.308
Chad	0.190	0.185	0.174
Guinea	0.272	0.339	0.326
Mali	0.257	0.307	0.294
Niger	0.317	0.266	0.249

Note: The first column reports the marginal R2 of the empirical best predictor model, taken from Table 2. The second and third columns reports the average in and out of sample R2 from cross-validated OLS regressions across ten folds. The results indicate that the in and out of sample R-square values are very similar in both training and test datasets in all countries.

Appendix D: EBP implementation

As mentioned in Section 3, we compute an analytic version of the EBP estimator. Under the analytic approach we calculate expected values of headcount poverty instead of using Monte-Carlo draws from the distributions of the error terms. The main advantage of using the analytic version of the EBP estimator is computational speed, To check that the results are not qualitatively different, we compare the two approaches for SAE in Burkina Faso. Table D1 reports the results of this comparison. The results show that the estimates under the two approaches are not meaningfully different.

Table D1: Comparison of the two approaches of implementing the EBP for small area estimation of headcount poverty in Burkina Faso.

Burkina Faso	Headcount poverty estimates	MSE estimates
Mean across areas (Analytic approach)	0.510	0.013
Mean across areas (Monte-Carlo approach)	0.509	0.013
Rank Correlation	0.997	0.902
Pearson Correlation	0.997	0.931
Mean Absolute Difference	0.009	0.002