# Small Area Estimation with Geospatial Data: A Primer

Joshua D. Merfeld[*] Haoyi Chen[†] Partha Lahiri[‡] David Newhouse[§]

**Abstract:** This paper provides a general overview and practical suggestions for using geospatial data in small area estimation (SAE). It discusses commonly used geospatial variables and publicly available repositories for this type of data. It then covers the basics of SAE methods as well as an introduction to the different types of data structures used with geospatial data, with a focus on the R programming language. Finally, it provides an in-depth example of how to pull geospatial data and use it to generate small area estimates using an accompanying repository on GitHub.

[*] KDI School of Public Policy and Management and IZA; merfeld@kdis.ac.kr
[†] UN Statistics Division
[‡] University of Maryland, College Park
[§] World Bank and IZA

## Table of Contents

# I. Introduction

In 2015, the United Nations General Assembly agreed to a new set of development objectives: the Sustainable Development Goals (SDGs). While poverty remains one of the most-discussed goals, the SDGs consist of 17 goals and 169 targets, encompassing a diverse set of objectives, like zero hunger, better access to health care and education, gender equality, and clean water, to name a few. Accurate monitoring of the targets and indicators under each SDG goal is critical to understand progress towards meeting the SDGs by 2030. For example, the first target for the poverty SDG is the elimination of extreme poverty by 2030. The only way to understand if countries are meeting this objective is to reliably estimate the proportion of persons in living in extreme poverty.

Historically, national statistical offices (NSOs) have relied on household probability sample surveys to accomplish this measurement. Yet, such household surveys suffer from several key drawbacks: large-scale surveys are expensive to implement, meaning countries face trade-offs between the number of households surveyed and the frequency with which NSOs implement nationally representative surveys. This leads to a very particular problem: while household surveys give reliable information about progress towards different SDGs at very aggregated levels – such as provinces, states, or districts – they cannot reliably estimate different indicators at more geographically disaggregated small areas such as subdistricts. This makes targeting interventions more difficult, since we do not necessarily know exactly which small areas are poorest. In addition, the lack of geographically granular data from traditional surveys can make it difficult to evaluate the effects of interventions that affect small areas.

The goal of this guide is to show how combining surveys with publicly available geospatial data using Small Area Estimation (henceforth SAE) can partially surmount some of the drawbacks of survey data and present more geographically disaggregated estimates of different SDG indicators at the subnational level. The paper discusses the availability of different types of geospatial data, as well as how one can statistically integrate geospatial features with household survey data.[5] We focus both on the auxiliary data itself as well as the nature of the survey (training) data in order to incorporate geospatial data. The inclusion of geospatial coordinates in survey data has become more common in recent years, making integration of geospatial data easier.[6] Our focus throughout the paper is on estimating outcomes in levels, not changes. This is because welfare changes are harder to estimate accurately with publicly available geospatial indicators than welfare levels, and the estimation of welfare changes using geospatial data remains an area of active research.

The following section gives a broad overview of small area estimation and how geospatial data can serve as auxiliary data. Poverty and wealth remain the main focus, following much of the literature (Jean et al, 2016 Yeh et al, 2020, Chi et al, 2021, Engstrom et al, 2022), but recent work has also explored the use of geospatial predictors for labor market outcomes (e.g. Merfeld et al., 2022) agricultural yields (e.g. Bellow and Lahiri, 2010, Erciulescu et al., 2019 and Lobell et al., 2020)), and population density (Wardrop et al, 2018, Engstrom et al, 2020), to name but a few. We pay particular attention to what we do not yet know, as the practice of combining survey and geospatial data remains in its infancy, despite an explosion of popularity in recent years. Section

[5] We mainly discuss the use of predefined geospatial features rather than convolutional neural network models directly trained to imagery (Jean et al, 2016, Yeh et al, 2020). This is partly because the latter requires a different set of skills, and because preliminary evidence suggests that the former performs better when using typical household surveys as training data (Ayush et al, 2020, Engstrom, Hersh, and Newhouse, 2022, Sohnesen and Fisker, 2022).

[6] These coordinates are often randomly offset to protect the privacy of survey respondents, but existing research suggests that random offsets of the coordinates detracts only slightly from the accuracy of the estimates (Van der Weide et al, 2022).

three reviews the availability of different types of publicly available indicators derived from geospatial data. Section 4 reviews estimation methods and models in greater detail, focusing on methods supported by well-documented and publicly available software packages.

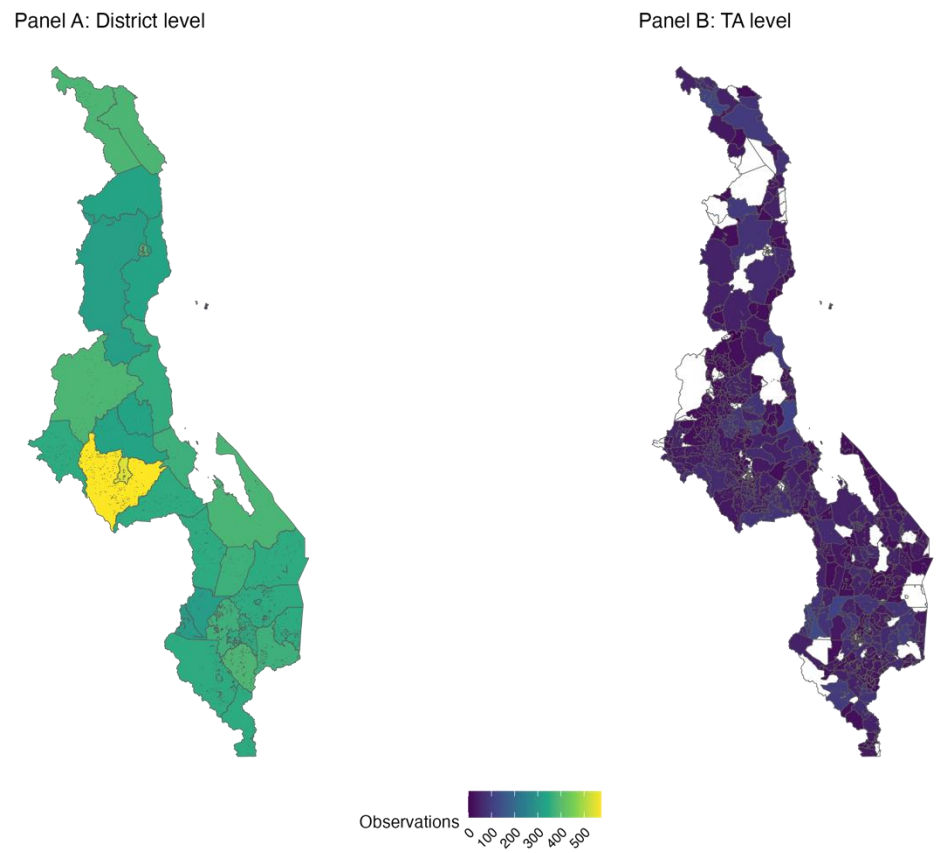
The final section discusses the skills required to be able to use geospatial data in small area estimation. For example, we discuss the different types of files common to geospatial data (e.g. rasters and shapefiles) as well as the common packages used in implementation. We focus mostly on the *R* language for statistical computing, given that its suite of packages allows one to implement the full workflow, from pulling geospatial data to implementing the estimation of different SAE models. Other statistics programs, like *Stata*, have packages for SAE.[7] However, they lack well developed packages for geospatial data. On the other hand, geospatial-specific programs, like *QGIS*, allow users to clean the geospatial data but do not easily allow for estimation of the SAE models themselves. We provide explicit recommendations for different *R* packages and offer examples, including in a companion GitHub repository with a detailed example. However, R packages are constantly evolving and new and updated packages will likely supersede existing tools. Even during the drafting of this document, several new packages related to geospatial data have entered the *R* ecosystem.

## II. Using geospatial data for small area estimation: The broader context

[7] The Stata SAE package estimates empirical best models, and the R povmap package also contains a Stata .ado and .hlp file that can run the package from within Stata.

SAE is a branch of statistics that has been actively studied for more than forty years (Ghosh, 2020). While there are different ways to implement SAE, all methods use either implicit or explicit models that combine survey data with more comprehensive auxiliary data, covering a larger share of geographic areas or population than groups surveys. The auxiliary data augments the survey data, which measures well-established concepts related to SDG indicators such as poverty, which are not typically available in the auxiliary data source. Augmenting surveys with auxiliary data can improve estimates where survey samples are too small and generate estimates when samples are unavailable. In other words, the models enable surveys to "borrow strength" from more comprehensive auxiliary data. If the model relationships are similar in areas with and without survey data, we can use the auxiliary data to impute synthetic estimates where survey data is missing or improve reliability by combining these synthetic estimates with survey data in areas with both. Similarly, SAE can be used to improve estimates for specific populations – like those with disabilities or ethnic minorities – even if survey sample sizes are small, either overall or for specific subgroups.

Figure 1: Malawi IHS5




The colors correspond to the number of household observations at the district (left panel) and traditional authority (right panel) for the 2019 Fifth Integrated Household Survey in Malawi.

An example can help fix ideas. Consider the case of Malawi, which has a nationally representative household survey from 2019, the Fifth Integrated Household Survey (IHS5).[8] The left panel of Figure 1 plots the number of observations per district. The IHS5 is considered to be representative at the district level, which means that the National Statistics Office feels the traditional survey-weighted direct estimates at this level are sufficiently reliable to warrant publication. This makes sense since the minimum sample size across districts is quite high. However, it is also clear from the map that districts can be relatively large, geographically, obscuring important heterogeneity in poverty rates within districts.

[8] https://microdata.worldbank.org/index.php/catalog/3818

It would, therefore, be useful to obtain reliable estimates one level below the district, the Traditional Authority (TA). In Malawi, there are 420 TAs, but only 32 districts. The right panel of Figure 1 shows very clearly that the household sample survey data cannot generate reliable estimates at the TA level. Many TAs have no survey observations at all (the white areas), while TAs that do have observations have relatively few; 121 TAs have no observations and, of those that do, the median is just 32 households. A direct survey-weighted poverty estimate based on 32 households is insufficiently precise to publish according to many countries' statistical standards.[9]

Instead of using the direct survey-weighted estimates, SAE allows us to integrate auxiliary data to improve reliability at the TA level, even where overall samples are small or non-existent. The main benefit of using this auxiliary data is that it is collected even in places where the survey data does not collect data. Traditionally, SAE practitioners have used administrative data of different types or censuses to serve as auxiliary data. For example, the US Census Bureau creates small area estimates of poverty using census data as well as federal income tax returns, among other administrative data. In addition, the World Bank has supported the production of several poverty maps in developing countries based on census data, generally following the method articulated in Elbers, Lanjouw, and Lanjouw (2003). For Malawi, there was a census conducted in 2018, just one year after the IHS5. This census includes detailed information from each household; in particular, it includes information on household assets, education levels, and household size and

[9] If the underlying poverty rate is 50 percent, the variance of the estimated poverty rate is approximately 0.0078 and the standard error is approximately 0.088, implying a coefficient of variation of approximately 17.7. Country thresholds for uncertainty vary. They can be as high as a median CV of 0.61 for the US American Community Survey, but are typically much lower, between 0.15 and 0.25. However, the American Community Survey will also suppress any statistics based on fewer than 50 observations.

age composition. Importantly, the IHS5 includes many of the same indicators. Continuing with our example of poverty, SAE allows us to model the relationship between the welfare variable (or poverty status) and these household characteristics in the IHS5 and then utilize both the census and survey data to estimate poverty at the TA level throughout the country, based on this estimated relationship.

A sound source of auxiliary data source is critical for SAE, yet countries often encounter challenges when trying to access high-quality auxiliary data. Access issues arise when NSOs do not have a recent census to use, or struggle to obtain data from other government departments, which may require governance agreements and protocols for data sharing, in order to ensure consistent updates for SAE models. Obtaining high-quality auxiliary data, encompassing factors like coverage, accuracy, timeliness, update frequency, and the presence of predictive auxiliary variables, is often difficult in countries with less developed statistical systems.

Within this context, the use of SAE with geospatial data can significantly improve on traditional direct survey-weighted estimators. Geospatial auxiliary data is a second-best option compared to recent high-quality household census or administrative data, but as noted above, the latter are often not available. In a variety of geographic contexts, combining survey and geospatial data has increased the precision of poverty estimates relative to direct survey estimates by an amount roughly equivalent to increasing the size of the sample by a factor of 2.5 to 6, depending on the context and indicator (Masaki et al., 2022; Newhouse et al., 2023; Merfeld et al., 2023). Small area estimation, regardless of the source of auxiliary data, introduces model error. In applications that have tested the use of SAE with geospatial data, the benefits of including predictive information

from additional areas tends to outweigh the inaccuracies created by model error. As a result, small area estimates are usually more reliable than direct survey estimates, especially for poverty estimation, although this is not always true and will vary depending on the indicator, the geographic context, the quality of the geospatial auxiliary data, and modeling strategy. In addition, small area estimates are substantially more precise than direct survey-weighted estimates, and can in some cases enable reporting at more disaggregated administrative levels.

Several statistical methods have been used for SAE in the past. Elbers et al. (2003) – through their publication and work at the World Bank – popularized the use of a unit-level model to link census data with sample survey data for small area estimation of poverty in developing country contexts. However, the use of SAE for poverty estimation has a much longer history. Small area poverty estimation dates back to at least the well-celebrated paper by Fay and Herriot (1979), who also proposed a two-level model of transformed direct survey-weighted estimates to estimate income for small places using the 1970 Census of Population and Housing in the United States.[10] While Fay and Herriot used an empirical Bayes method for their application, subsequent researchers have also considered pure Bayesian implementations (Morris and Tang, 2013, Arima et al (2017) , Hirose and Lahiri (2020). More similar to Elbers et al (2003) are the families of unit-level models. Like for area-level models, these can be implemented either as Empirical Bayes/Best (EB) models (see Ghosh and Lahiri, 1987, Battese, Harter, and Fuller, 1988, and Molina and Rao, 2010) or purely Bayesian methods (Ghosh and Lahiri, 1992, Hall and Maiti, 2006). It is also possible to employ time series models to combine information over time (see, e.g., Pfeffermann and Tiller (2006), time series cross-sectional models to combine information over time and areas (e.g., Datta

[10] This in turn built on earlier applications of small area estimation such as Carter and Rolph (1974) and Efron and Morris (1977)

et al 1999) , and spatial models to incorporate spatial correlations (e.g. Vogt et al., 2023). Bayesian Structural Equation Modeling has also been applied in many cases (Gething et al, 2015, Steele et al, 2017, Van der Weide et al, 2022), including by the innovative spatial data repository maintained by the Demographic and Health Survey program. These models often include parameters that smooth changes over space, which distinguish them from standard Empirical Best estimates. It is not clear whether these smoothing parameters increase or decrease accuracy in practical contexts.

Recent advancements in machine learning have also opened up new avenues for SAE, including the direct use of tree-based machine learning methods (e.g. Chi et al, 2022, Merfeld and Newhouse, 2023) as well as the integration of random forests with mixed effects for SAE (Krennmair and Schmid, 2022). More sophisticated machine learning methods such as neural networks (Yeh et al, 2022) or even large language model transformer models (Manvi et al, 2022) can also be employed. However, these machine learning based SAE methods need more theoretical and empirical investigation. Moreover, they require specialized skills and knowledge to deploy with confidence.

Broadly speaking, many of the methods mentioned in the preceding paragraphs differ along three key dimensions. The first is the level at which models are proposed. For example, in the case of Malawi, when using geospatial data there are three possible modeling strategies to estimate poverty at the TA level. The first is to model direct survey-weighted estimates of poverty rates across TAs. This has its roots in the method suggested by Fay and Herriot (1979) and is often referred to as an "area-level" model, where "area" refers to the target domain at which the indicator is estimated. A second possible modeling strategy is to model aggregate data at a lower level of aggregation, which we refer to as a "sub-area model". In Malawi, this could be the enumeration

area, for example, of which there are 18,700, compared to just 436 TAs. Using this modeling strategy, we can estimate enumeration-area-level poverty and then take a weighted mean to aggregate these up to the TA level, weighting by the population of each EA. Finally, one can propose a model at the household level. Under this modeling approach, a welfare variable is imputed for individual households, and then converted to poverty rates and aggregated up to the desired target area (TAs). The household-level and subarea-level models are preferable when reliable auxiliary data is available at lower levels than the area level, which is typically the case with geospatial data. This is because they estimate the model at the lower, more disaggregated, level before aggregating the predictions to the target area level.

The second key difference is the specific estimation method used, given an assumed model. In the Bayesian and Empirical Bayesian approach the small area parameters of interest are conditioned on the sample data.[11] In other words, estimation is based on the conditional distribution of an area-specific random intercept given the observed data, under the assumed model. In contrast, synthetic estimation such as Elbers et al (2003) and newer machine learning methods are based on purely synthetic predictions, using unconditional distributions. For non-sampled areas, this distinction is irrelevant, because there is no sample data for the Bayesian or Empirical Bayesian methods to condition on. For in-sample areas, the Bayesian, empirical Bayesian or Empirical Best Predictor (EBP) methods, unlike purely synthetic methods, are a combination of the direct survey-weighted estimates and synthetic model prediction. These methods typically give more importance to the sample survey data, vis-à-vis the synthetic prediction, when the sample is large and/or synthetic predictions are less precise. Conversely, less weight is given to the sample survey data when the

[11] Empirical Bayesian approach is also sometimes referred to as “Empirical Best” approach.

sample is small and/or the synthetic predictions are precise. Conditioning on survey data is a desirable feature in small area estimation, especially when using data that is linked to the survey at a sub-area level like a village or enumeration area, as is typically the case with geospatial data. Because this is a more aggregate level than the household, the synthetic predictions are less precise, and greater weight is typically given to the survey when using sub-area geospatial indicators as auxiliary data than when using a household census as auxiliary data.

Until recently, the emphasis on Bayesian or empirical Bayesian models among statisticians has restricted the specification and the class of estimators that have been considered for small area estimation. However, synthetic methods, including recently developed pure machine learning methods, can benefit from more flexible model specifications that better handle non-linearities and interactions among predictors. Better understanding the trade-offs between the flexibility of pure machine learning model and the benefits of conditioning on sample data is an active area of research, as is integrating machine learning approaches into Bayesian or Empirical Bayesian models that condition on the sample data.

The third key difference in small area estimation methods – and the focus of the rest of this guide – is the type of auxiliary data used. Traditionally, practitioners have used detailed census or administrative data as auxiliary data. However, this type of data is not always available, especially in less developed countries.[12] Battese, Harter, and Fuller (1988) pioneered the use of satellite-derived imagery in small area estimation, proposing a method to estimate the area of different crops in a small part of the United States. More recently, a large literature has expanded on the use

[12] The UN's toolkit for SAE summarizes many of these issues. You can find it here (https://unstats.un.org/wiki/display/SAE4SDG/).

of remote sensing data for estimating poverty (e.g. Bellow and Lahiri, 2010, Jean et al, 2016, Yeh et al, 2019, Chi et al, 2021, Engstrom et al, 2021). An important advantage of using satellite-derived data is its availability: satellite data is generally available across the entire globe, meaning that data-poor areas are covered by satellites, even if they lack more detailed census data. While somewhat less predictive, geospatial data is much more widely available – both geographically and temporally – for national statistical offices and researchers.

## III. Geospatial data availability

### A. Geospatial Indicators

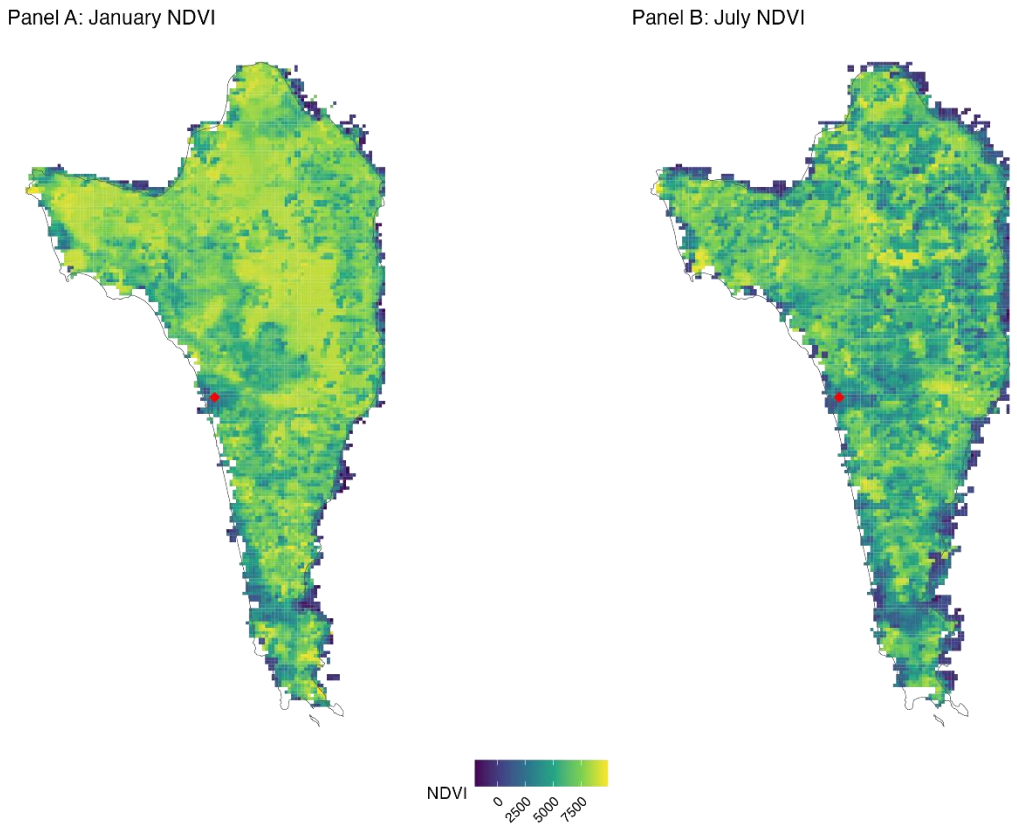
In recent years, the availability of geospatial data has exploded. There are several publicly available repositories of this type of data, with perhaps the most popular and well-known at this time being Google Earth Engine (GEE).[13] GEE offers hundreds of different datasets, many of which span up to three decades. GEE does not create the data, *per se*; instead, it is a repository of externally, publicly available data from across the globe. We focus here on just a few of the datasets available on GEE, but we encourage readers to explore the wide offerings available on the GEE website.

As an example, consider NASA's Moderate Resolution Imaging Spectroradiometer instrument, more commonly referred to as MODIS, which is deployed on two separate satellites.[14] NASA makes these data publicly available and multiple different data products on GEE are derived from

[13] https://earthengine.google.com/. Microsoft's Planetary Computer is a similar product that makes geospatial data publicly available and can be accessed at https://planetarycomputer.microsoft.com/
[14] https://modis.gsfc.nasa.gov/data/

the underlying imagery. One commonly used variable is the Normalized Difference Vegetation Index, or NDVI. NDVI is an estimate of how "green" a given area is and can be used to identify different types of land (e.g. waterbodies, deserts) and, where there is vegetation, how dense the vegetation is.

Figure 2: NDVI for Phu Quoc, Vietnam (2022)




NDVI is from MODIS, pulled from Google Earth Engine. Values for both maps are from 2022. The red mark on the left side of each map marks the location of the largest town/urban area on the island. The raw NDVI data is scaled up by 10,000.

Some variables vary substantially across the year due to seasonality. Continuing with our NDVI example, consider the island of Phu Quoc, Vietnam, in Figure 2. The maps contain NDVI values for two separate points in time: January of 2022 (left panel) and July of 2022 (right panel). The scales are identical across the two maps, showing the large differences in NDVI values, particularly in the middle of the island. This is because NDVI tends to vary greatly throughout the year,

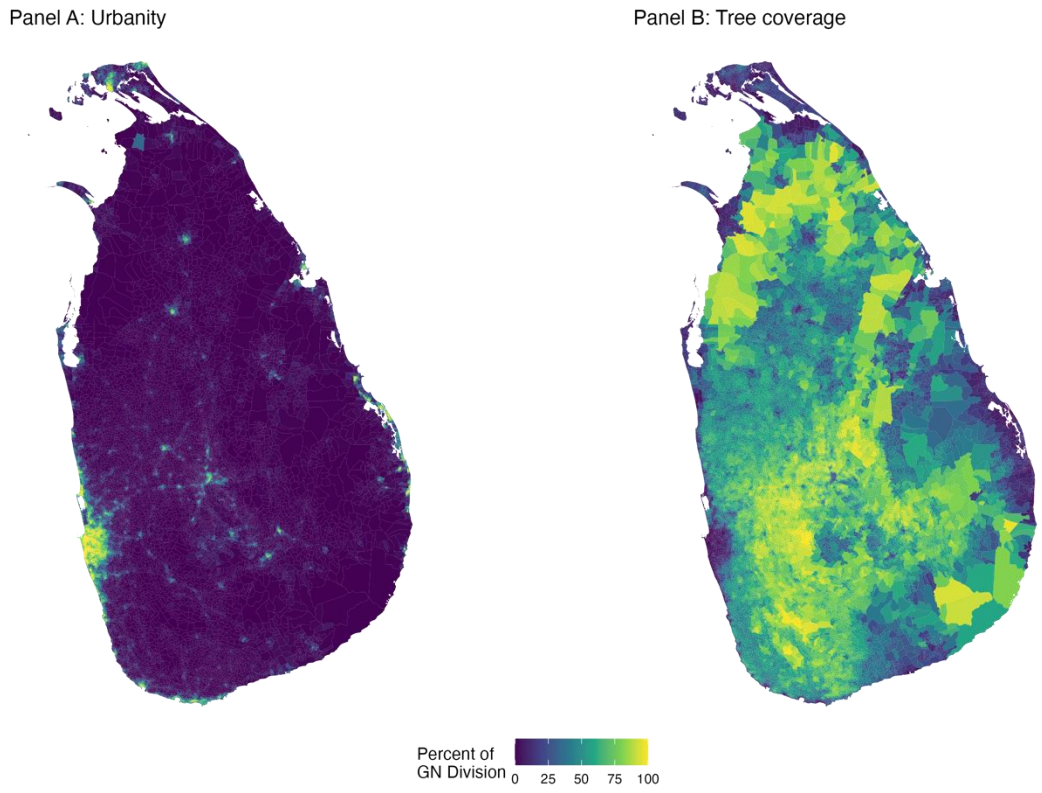
meaning the use of "mean" NDVI for the year – or at just one point in time – can be misleading. Therefore, we recommend matching satellite indicators to the specific month of interview data collection, if possible, particularly if one expects this seasonality to be important. In addition, it may be worthwhile to include several different possible variations of NDVI, like mean, minimum, maximum, and standard deviation throughout the year, as well as lagged values, when appropriate. We return to this point below.

The resolution of the satellite data is also important. In figure 2, the NDVI colors are formed by small grid cells, each of which has exactly one color. In the case of the MODIS NDVI shown here, each grid cell is approximately 250m by 250m. The length of each edge is referred to as the "resolution" of the variable. Resolution differs by indicator, depending on the satellite and instrument that were used to capture the image it was derived from, and can be as small as several centimeters or as large as many kilometers. For example, TerraClimate has estimated monthly temperature across the entire globe from 1958 until present. [15] Many of the temperatures are estimated; since actual temperature observations are unavailable in many places, a model is used to impute them. As a result, the resolution of monthly temperature is 4 km, which is somewhat higher than NDVI, which is directly measured. It is important to keep resolution in mind when thinking about candidate geospatial variables, since variables with much lower resolution (larger grid cells) will tend to vary less across space and may be less predictive of the variance in key outcomes when the domains of interest are geographically small.

[15] https://www.climatologylab.org/terraclimate.html

Land classification is another variable on Google Earth Engine that is often predictive of development outcomes. This is partly because land classification is strongly correlated with population density, or how urban a place is, which in turn is strongly correlated with many development outcomes such as poverty. GEE provides estimated land classification from Copernicus, which has a resolution of 100m.[16] The land classification layer classifies each grid cell based on observable characteristics into different land classification categories, including forests and built-up land. One way we use these land classifications in our work is to calculate the percentage of a given administrative area that is made up of different land classifications.

Figure 3: Land classification percentages by GN Division, Sri Lanka




The left map shows the percentage of each GN Division in Sri Lanka that is considered urban according to the Copernicus land classification variable. The right map shows the percentage of each GN Division that is covered by trees.

[16] https://lcviewer.vito.be/2015

Figure 3 shows an example of this for Sri Lanka. The left map shows the percentage of each Sri Lankan GN Division that is estimated to be urban, according to Copernicus' land classifications, while the right map shows the percentage of each GN Division that is estimated to be covered in trees. Since rural and more remote areas tend to be poorer than urban areas, it is perhaps unsurprising that land classification is generally a strong predictor of many outcomes. From both maps, it is immediately clear where the large urban centers are around the coast, with the large urban center of Colombo jumping out in the west-southwest portion of the map. Estimated urbanity there is high, while tree coverage is quite low, consistent with what we would expect from large urban agglomerations.

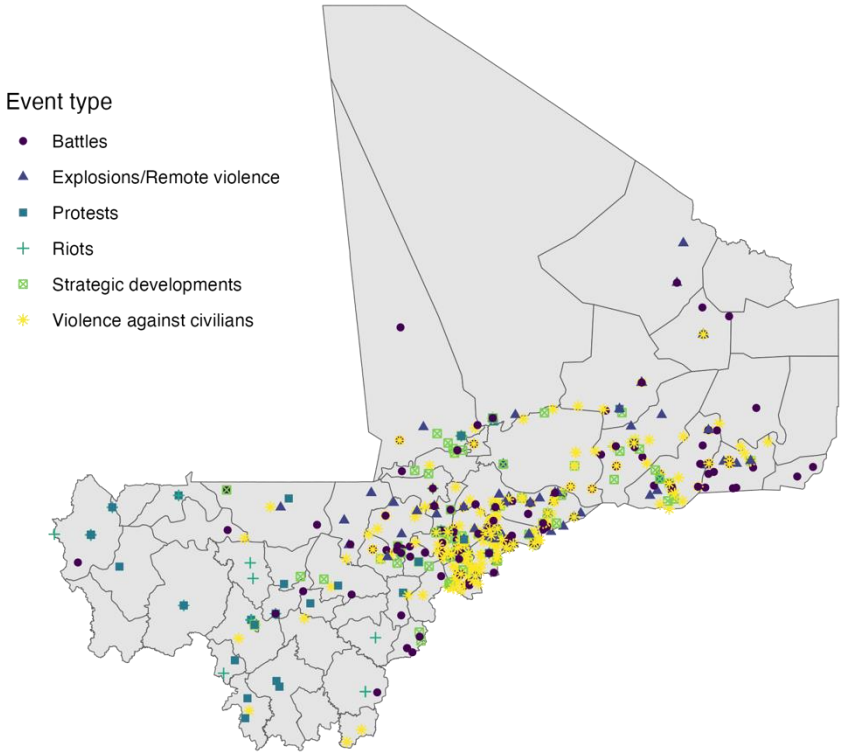
There are many other useful and frequently used data layers in GEE – including nightlights and pollution estimates – but this section now turns to other geospatial indicators not available in GEE. One example is ACLED, the Armed Conflict Location & Event Data Project.[17] ACLED collects data on many sorts of violent encounters and protests across the world, including estimates on fatalities. These events are geolocated, with GPS coordinates available that allow us to pinpoint approximate locations and match these to survey data using shapefiles.

In some countries, the number of violent encounters in different areas can be a strong predictor of many development outcomes. Consider the case of Mali, for example, which suffers from terrorist attacks and has had several coup d'état over the last decade.

Figure 4 shows the location of different violent events in Mali in 2019. ACLED provides six different event types, which are color- and shape-coded in the map.

[17] https://acleddata.com/

Figure 4: Violent events in Mali, 2019




The map shows the location of different violent events in Mali in 2019, according to ACLED.

How can ACLED data be included as geospatial predictors in SAE? There are many different options, and the best choice depends on the context. For example, if the outcome of interest is something like assets – which tend to vary less across time than monetary welfare measures – the total number of events or fatalities of different types in the administrative region in the previous year may be a useful predictor. On the other hand, if the outcome is something that may respond immediately to violence – like school attendance – then perhaps using the number of violent events around the time of survey enumeration would be a good predictor. This example should make clear that the optimal way to include geospatial covariates as predictors can vary across contexts. Therefore, knowledge about the outcome of interest and the survey enumeration strategy is

important to maximizing the predictive power of these covariates; matching monthly level violent conflicts to survey data may be less predictive of outcomes that vary less over time – like human capital, e.g. educational attainment (as opposed to attendance) – whereas using longer term averages of conflicts may be more predictive in these cases.

Other commonly used geospatial covariates include things like climate and weather. There are several options, including the aforementioned TerraClimate. Simple means of precipitation/temperature across different regions of the country are one possible way to include these variables, as are contemporaneous rainfall, using something like CHIRPS (Climate Hazard group InfraRed Precipitation with Station) data.[18] Other options include variables like degree days – the number of days in which the temperature rises above a certain value (e.g. 28 degrees Celsius) – and heatwaves, either as averages or as contemporaneous values. When modelling agricultural yield, for example, contemporaneous values are often quite predictive of productivity, especially in predominantly rainfed areas, like Sub-Saharan Africa.

Distance to different facilities can also be correlated with development outcomes. For example, in ongoing work to estimate disaggregated prices, we calculate the distance to different provincial capitals from each administrative area and include these distances as candidate predictors, as well as the distance to other markets. Because we are not aware of a public dataset with this information, we input the location of these capitals by hand, using GPS coordinates.[19] We have also previously included distances to ocean, nearest port, or other countries, as well. Recently, the FAO has created

[18] https://www.chc.ucsb.edu/data/chirps

[19] AI tools such as ChatGPT can be quite helpful for this type of exercise. However, we caution users to always double check answers given by AI tools, as they are prone to giving incorrect information in some situations.

a dataset called "Rivers of Africa" that plots the location of rivers of different types throughout the continent. We can calculate the distance to different rivers or the total length of different types of rivers within a given administrative area.

One option that provides excellent data for some countries is OpenStreetMap (OSM).[20] OSM provides information about roads and numerous types of amenities. OSM data, however, is self-reported by an open-source community. It may therefore not always be complete in many less developed countries or in less urban areas (Barrington-Leigh and Millard-Ball, 2019). As such, users should be wary of this, as missing data in parts of a country can lead to biased predictions. For example, information on medical clinics in a given area may not be reported. If we assume there are no medical clinics – which is how this might show up in the raw data – when there really are, this can lead to biased estimates if this mismeasurement is not random, but rather systematic.

There are also datasets on building footprints that are derived from satellite data. Microsoft,[21] Google,[22] and the World Settlement Footprint (WSF)[23] have released layers covering many countries. Worldpop has also released publicly available summary statistics for several countries in Sub-Saharan Africa.[24] These building footprints can be aggregated up to administrative areas in different ways, providing yet another possible covariate to use in small area estimation. Unfortunately, however, they are not updated regularly, at least as of this writing. As with OSM, these footprints are not comprehensive in all areas and the cited confidence of predictions in some

[20] https://planet.openstreetmap.org/
[21] https://www.microsoft.com/en-us/maps/bing-maps/building-footprints
[22] https://sites.research.google/open-buildings/
[23] https://geoservice.dlr.de/web/maps/eoc:wsf2019
[24] https://wopr.worldpop.org/?/Buildings

areas is much lower than in others. Since this data is satellite derived, some of these errors may be due to cloud cover. In these cases, it *may* be less systematic than measurement error in OSM, depending on the relationship between cloud coverage and outcomes of interest.

Not all candidate variables are geospatial variables, per se. WorldPop[25] provides estimated population counts across space and time for many countries of the world. They also provide estimates for different demographic groups and across age and gender. These are constructed both using "bottom-up" methods from survey or microcensus data, and "top-down" methods that allocate reported census statistics based on building layers. If these estimates are based on older census data, they may be dated. WorldPop provides these data in a geospatial format, however, allowing users to match population estimates to survey data. We return to this point below.

Another possible source of auxiliary information comes from cell phones and mobile data. OpenCellID offers locations of cell towers in many areas across the globe.[26] More detailed cell data records (CDR) are sometimes available and can be quite predictive of important outcomes. Aiken et al. (2022), for example, show how leveraging CDR and machine learning methods can help governments target the poorest for humanitarian assistance. However, CDR data is highly confidential and, as such, access is difficult. In addition, CDR data necessarily covers only individuals with mobile phones – and in many contexts, only mobile phones on the network of a specific operator – meaning they do not necessarily represent the entire population. This can be particularly problematic when trying to impute outcomes in areas with no survey data.

[25] https://www.worldpop.org/
[26] https://www.opencellid.org/downloads.php

The rapid improvements in machine learning with imagery has led to an explosion of open access data based on this imagery. The most recent example is MOSAIKS, which offers a set of variables derived from satellite imaging and machine learning.[27] These variables are generally not directly interpretable, but they cover large portions (of populated areas) of the globe and can be predictive of many outcomes (Rolf et al., 2021). If practitioners want to be able to interpret the relationships between outcomes and geospatial covariates, then MOSAIKS may not be the best option. However, when the model's predictive accuracy is the main concern, adding MOSAIKS features can be helpful.

Similar to MOSAIKS are the spatial features computed by the Sp.feas package in Python.[28] This package computes a variety of contextual spatial features, calculated by comparing pixels to their neighbors, that are predictive of building density, population density, poverty, and wealth (Chao et al, 2021; Hersh et al, 2021; Engstrom et al, 2022).

Recently, Chi et al. (2022) released estimates of wealth indices at relatively fine resolutions for most developing countries. Some of the predictor variables were proprietary connectivity data from Meta. However, they are available in formats common to geospatial variables, like raster files, which we cover below.

Finally, we want to mention one last issue: masking. Masking is the process of removing certain areas from the data. Perhaps the most common type of masking is removing areas with zero population, although it is also common to mask based on other variables, like using only cropland,

[27] https://www.mosaiks.org/
[28] https://jgrss.github.io/spfeas

depending on the context and goal. The use of masking is an ongoing research topic. However, in our own work, we have found that it rarely matters, mostly because areas of zero population will 1) not have any survey data and 2) will not be given any weight when we aggregate predictions using small area estimation methods. However, masking may provide more informative geospatial measures that are more predictive of outcomes.

### B. Linking geospatial data to survey data

While there is now a plethora of different options available for geospatial data, a key requirement remains: training data. In this context, training data usually refers to household survey data that includes the key outcome of interest. For example, if we are interested in using SAE to improve estimates of poverty, we need survey data with a measure of welfare, defined as household income or consumption adjusted for household composition. In traditional SAE using census data, the training data requires additional information on proxy correlates of welfare, which is used to estimate a model that can then be applied to the census. For geospatial small area estimation, only location data is required in order to merge the geospatial data into the survey data.

The resolution at which identifying location information is available is also important. Having the geocoordinates of household residences is ideal, but is rarely available due to confidentiality restrictions, though this is not always true for NSOs themselves. More commonly obtainable are geocoordinates at the enumeration area level, which are often jittered by adding a random offset to preserve confidentiality by preventing the identification of specific villages. Other times, as discussed below, survey data is linked to geospatial data based on administrative identifiers. In that case, it is best if identifiers can be obtained at a highly disaggregated level such as a village. This enables geospatial data to be extracted at a granular level. Linking more disaggregated

geographic data generates more precise and accurate estimates and can also be useful to detect and partially correct for errors due to a non-representative sample. A less desirable but still feasible solution is to use survey identifiers at the target area level. Geospatial data can then be incorporated into area-level small area estimation models. However, area-level models based on geospatial variables are usually less accurate and precise, and we advise that it is best to link geospatial data at the lowest level possible.

The size and nature of the training data are also important. Particularly when using more flexible machine learning methods to estimate models, more training data can be used to build richer models that yield more accurate estimates. Supplementing household survey data with partial registries that collect proxy outcomes from households can greatly improve the accuracy of welfare estimation based on geospatial variables (Gualavisi and Newhouse, 2022).

A shapefile is the key tool that can help integrate survey and geospatial data in many contexts. A shapefile is essentially a list of each administrative area and its corresponding geography expressed in georeferenced polygons. All the maps in the previous sections were created with shapefiles. How do we integrate the survey data with a shapefile? There are two possibilities. The first is for the shapefile and the survey data to include the same *identifiers*. In Malawi, we can match households to the shapefile using the Enumeration Area identifier, since they are consistent across the two data sources. However, this is not always the case; sometimes identifiers change across years, for example. In other cases, the lowest level of aggregation may not be low enough to allow for more accurate SAE methods. For example, the Indian National Sample Survey (NSS) only has district-level identifiers, meaning that linking geospatial data at more disaggregated levels for SAE

is not possible. However, area-level models could in theory be used to improve the accuracy and precision of the district-level estimates from the sample.

The second option for linking survey and geospatial data is using GPS coordinates. Nowadays, it is quite common for datasets to include geographic coordinates for enumeration areas. To protect the identity of respondents, these geographic coordinates are often randomly offset (jittered) by a small amount – usually between one and five kilometers – but generally allow for acceptable matching with a shapefile. In fact, recent research indicates that random jittering has little impact for poverty and wealth mapping (Burke et al, 2021, Van der Weide et al., 2022). Thankfully, the list of household surveys with jittered GPS coordinates is large and growing. Many newer DHS surveys include GPS coordinates, as do most of the World Bank-backed LSMS surveys. The inclusion of jittered coordinates in DHS data has enabled the publication of modeled estimates of outcomes at the raster level from selected DHS surveys in the DHS spatial repository. We encourage statistics offices considering integrating geospatial data into survey data to collect the most granular geographic information possible, while still protecting the identity of survey respondents. Jittered GPS coordinates are one way to do this.

## IV. Geospatial SAE methods

### A. Alternative models

This section turns to different analytical models, methods, and software packages that can be used for small area estimation with geospatial data. This includes the selection of predictors, transformations, models, and generating point and uncertainty estimates. While much of the discussion is relevant to small area estimation in general, we focus on cases where geospatial data

is used as auxiliary data. An important caveat with any discussion of alternative methodologies for SAE is that the relative performance of different methods depends on the underlying structure of the sample and auxiliary data, in ways that are not yet fully understood. Because different models can perform better or worse in different contexts, general claims regarding the superiority of estimates generated from one method or model compared with others should be treated with appropriate skepticism. We therefore present a variety of methods which have been shown to work well in different applications and are supported by well-documented software packages.

To help fix ideas, we return to the example generating small area poverty estimates in Malawi. The household survey is considered to be representative for the 32 districts in the country, but we are interested in estimating poverty or wealth indicators for the 420 Traditional Authorities, which comprise the "target area". There is also an administrative unit below that, the enumeration area (EA), of which there are 18,700. Geospatial data can be linked to some survey or census data at the EA level, using available shapefiles. We refer to EAs in this context as the subarea, which is the lowest administrative level at which geospatial data can be linked to survey data.

We start by considering common small area models that can be estimated using existing software packages. One early model in this class was proposed by Fay and Herriot (Fay and Herriot, 1979) and can be described as follows:

$\hat{P}_a = \beta_1 X_a + \eta_a + \varepsilon_a,$

where $\hat{P}_a$ represents the survey-weighted poverty rate in area a, $X_a$ is a vector of predictor variables aggregated to the area level, $\eta_a$ is a random effect specified at the area level a, and $\varepsilon_a$ is the sampling error of $\hat{P}_a$. The random effects and sampling errors are assumed to be independent

normally distributed with zero means and unknown random effect variance $\sigma_\eta^2$ and known $\sigma_{\varepsilon i}^2$, respectively. Like most small area estimation models currently used, this type of model assumes spatial homogeneity, in the sense that both on the vector of unknown regression coefficients $\beta_1$and the random effects variance component $\sigma_\eta^2$are assumed to be the same for all small areas in the sample. The area-level poverty estimates can be estimated directly, as is done by the Small-Area Income and Poverty Estimates (SAIPE) program in the US Census Bureau. It is also common to take transformations, such as the arcsin transformation, which can facilitate variance smoothing (see, e.g., Fay and Herriott, 1979) and ensure the small area estimates are within the admissible range (see, e.g., Casas-Cordero et al. 2016).

The Fay-Herriot model has spawned a large literature and has been a workhorse model for small area estimation for decades. It is implemented with several useful estimation options in the povmap and EMDI packages in R, as well as the fayharriott (Halbmeier er al, 2019) and fhsae commands in Stata. However, area level models suffer from a key drawback in some SAE applications: They require that all data be aggregated to the target area level. For example, if we wanted to use nightlights as a predictor in an area level model, we would need to aggregate nightlights to the area level. This means we lose much of the possibly predictive heterogeneity in features that are available at lower levels. Since target areas are often much larger than sub-areas, including in the Mexican context, we recommend the use of auxiliary data at the lowest available level possible.

In the Fay-Herriot model, the sampling variances are assumed to be known. But in practice, they are not known and need to be estimated. Best practice is to estimated these area variances using variance smoothing techniques (such as the ones proposed by Fay and Herriot (1979), Otto and

Bell (1995), Ha et al. (2014), Bell et al. (2007), Hawala and Lahiri (2018), You, 2021). This is a non-trivial exercise and has not yet been implemented, as far as we know, in any of the available software packages. Another issue with such modeling is that measures of uncertainty of EBP such as MSE estimates fail to incorporate the additional variability incurred due to estimation of sampling variances because the sampling variances are, again, assumed to be known.

A final concern with area-level models is that the random effect variance estimates, when using standard estimation techniques such as maximum likelihood and residual likelihood, may be zero. This is particularly likely to occur when the variance of the outcome is close to zero, for example because sample poverty is close to zero or one in a particular area. This can cause an overshrinkage problem, in the sense that the estimate will come purely from the synthetic prediction without giving any weight to the direct sample estimate. A solution to overcome this problem was proposed by Li and Lahiri (2010) (also see Yashimori and Lahiri, 2014) and subsequently used by Casas-Cordero et al. (2016) in implementing an area-level EBP-based method for small area estimation of poverty in Chile. This solution has been implemented in the emdi and povmap packages in R, as well as the Fayherriot package in Stata, and is recommended to avoid cases where the estimated variance of the random effect is zero.

Two other classes of models are a preferred alternative to area-level models when recent population census data are not available and auxiliary data are available at a level below the target area. These models, unlike the area level model, can utilize the auxiliary data at the more disaggregated sub-area level, leading to more precise and accurate estimates. The first is referred to as a “sub-area model”, where outcomes are modeled at the subarea (EAs in Malawi, for example,

or AGEBs in Mexico) level and then predictions are aggregated up to the area level. This method is the most common in small area estimation using geospatial data and is the method used in Merfeld et al. (2022).

Such a model can be written as follows:

$G(\hat{P}_{as}) = \beta_1 X_{as} + \beta_2 X_a + \eta_a + \varepsilon_{as}$,

where $G(\hat{P}_{as})$ is a transformation of survey-weighted proportion $\hat{P}_{as}$ ; $X_{as}$ is a vector known sub-area level predictor variables; $X_a$ is a vector of known area level predictor variables; $\eta_a$ is an area random effect while $\varepsilon_{as}$ is a sub-area specific sampling error term. The error terns $\eta_a$ and $\varepsilon_{as}$ are assumed to be independently normally distributed with zero means, zero covariance, unknown random effect variance $\sigma_\eta^2$ and known sampling variances $\sigma_{\varepsilon as}^2$. The spatial homogeneity assumption applies both to the vector of unknown regression coefficients $\beta_1$and $\beta_2$and the random effects variance component $\sigma_\eta^2$, which are assumed to be the same for all areas. Since $\hat{P}_{as}$ is a proportion, the arcsin transformation implemented in the R povmap package may be suggested, which sets $G(\hat{P}_{as}) = \arcsin(\sqrt{\hat{P}_{as}})$.

A second class of model that can incorporate sub-area level predictors is sometimes referred to as a "unit-context model," where outcomes are modelled at the *household* level as a function of predictors at a more aggregate level (e.g. Lange et al, 2022, Masaki et al., 2022). This can be viewed as a special case on the unit-level model proposed by Battese et al. (1988), where household-level predictor variables are not used. Unit-context models are well-suited to predicting non-linear functions of underlying continuous variables, such as poverty or stunting rates. A one-fold unit unit-context model can be written as follows:

$$G(Y_{asi}) = \beta_1 X_{as} + \beta_2 X_a + \eta_a + \varepsilon_{asi},$$

where $Y_{asi}$ is a monetary welfare measure, namely household size-adjusted income or consumption, for household i living in sub-area s, within target area a; $X_{as}$ are sub-area level geospatial indicators; as before $X_a$ are the same variables aggregated to the target area level. The error terms $\eta_a$ and $\varepsilon_{asi}$ are assumed to be independently normally distributed with zero means, zero covariance, unknown random effect variance $\sigma_\eta^2$ and error variances $\sigma_\varepsilon^2$. Spatial homogeneity is assumed both for the vector of unknown regression coefficients $\beta_1$ and $\beta_2$ and the variance components $\sigma_\eta^2$ and $\sigma_\varepsilon^2$.

Meanwhile, a two-fold unit-context model adds a sub-area random effect to the model, and can be written as follows:

$$G(Y_{asi}) = \beta_1 X_{as} + \beta_2 X_a + \eta_a + v_{as} + \varepsilon_{asi},$$

where $v_{as}$ is a random intercept at the subarea level that, like $\eta_a$, is conditioned on the sample survey data.

The two-fold model has also been implemented in the R povmap package and the Stata SAE package, though the latter does not yet at the time of this writing allow for sample weights when estimating the two-fold model. When the survey is unbiased, the two-fold model tends to produce slightly more accurate point estimates, and more accurate estimates of uncertainty, than the one-fold model. The improvement in MSE estimates is larger when estimating unit-context models, where there is high correlation across sub-areas within areas because the predictor variables are identical for all households in a sub-area.

Does this model produce biased estimates? In practical applications, the true data generating process is unknown, making it impossible to assess the extent of bias in the parameters. In simulations where the true model is assumed to be a unit-level model, but a unit-context model is estimated, the use of a unit-context model leads to biased estimates of the variance components, primarily of the estimated variance of the household-specific error term $\varepsilon_{asi}$. Estimates of the regression coefficients $\beta_1$and $\beta_2$are not biased. In simulations, however, the impact on the bias of the point estimates produced under the unit-context model appears to be small.

When implementing a unit-context model, the population data will typically contain identical information for each household within each subarea. This is because the auxiliary data is linked to the household data at the subarea, ensuring that each household within that subarea has identical values of the predictor variables (e.g. a single night-time lights value applies to all households in each location). This commonly occurs when using geospatial data linked to the survey at the sub-area level. It can be tempting when implementing this model to drop duplicate records within each sub-area in the census data, such that the census data consists of one "representative household" per sub-area. Because population weights can be specified in many software packages, setting population weights equal to the estimated population of the sub-area would ensure proper weighting when aggregating in this case. Furthermore, dropping duplicate households in the census data greatly reduces computation costs, and generally has minor impacts on point estimates if the number of Monte Carlo replications is sufficiently large.[29] However, this procedure, if implemented naively, will increase estimates of Mean Squared Error. Therefore, when estimating

[29] For example, selecting L=100 replications will on average approximate each sub-area's headcount poverty rate to the nearest integer value. When specifying analytic rather than Monte Carlo calculations for point estimates Dropping duplicate observations has no impact on the point estimates.

Mean Squared Error for unit-context models, one should either include all units in the census data (including duplicates) or use the the MSE_pop_weights feature implemented in the R povmap package. This option instructs the algorithm to treat each observation in the census data as if contains duplicates when generating MSE estimates, which saves substantial computation time.

The majority of SAE models, including the ones proposed in three highly cited small area estimation papers – Efron and Morris (1975), Fay and Herriot (1979) and Battese et al. (1988) --- incorporate spatial heterogeneity only through area-specific intercepts. Therefore, the regression coefficients and variance components are assumed to be the same for all areas. This implies that that these models assume that the correlation between geospatial indicators such as night-time lights or urban extent and the outcome of interest is the same for all areas. In addition, the variance of the idiosyncratic error term is assumed to be the same for all areas, as is the variance of the area random effect before any shrinkage occurs. This is an important assumption, and it may not always be reasonable to assume that the same model parameters (e.g., slopes and sampling variances) apply for all areas when pooling a large number of small areas, even when the model includes area-specific random intercepts.

Because of the spatial homogeneity assumptions, determining which small areas should be estimated is an important problem. In particular, it may be useful to pool areas that are expected to be more homogenous with respect to slopes and sampling variances, when estimating SAE models. This is because spatially homogeneous models can perform poorly when there is a lot of heterogeneity in the underlying true model, such as cases where night-time lights are associated with larger welfare in one area but negatively in another (Tarozzi and Deaton, 2007).

Hobza and Morales (2013) relax the spatial homogeneity assumption by estimating a model that includes random area-specific slopes in addition to random area-specific intercepts. Lahiri and Salvati (2023) propose a flexible model that allows for area-specific slope coefficients, as well as heteroscedastic error terms. A data driven technique is used to generate small area estimates that are robust against model misspecifications. Both of these methods perform much better than standard EBP in settings in which there is a lot of heterogeneity in the true model across small areas. Unfortunately, however, the method has not yet been implemented in an easy-to-use and well-documented software package.

Another class of models uses more flexible machine learning methods instead of a linear model. One approach involves using machine learning methods such as extreme gradient boosting to predict the outcome directly (Chi et al., 2021; Mcbride et al., 2022; Merfeld et al., 2023). The second involves Mixed Effect Random Forest (MERF) models (Krennmair and Schmid, 2022), which allows for a conditional random effect in a random forest framework.

We briefly describe the machine learning based methods. In what follows, r denotes region that contains area a, and data from all regions can be used for small area estimation. Extreme gradient boosting methods can be written as follows:

$\hat{P}_{ras} = \sum_{m=1}^{M} f_m(X_{ras}, X_{ra}, X_r)$,

where each $f_m$ is a random forest prediction function; $\hat{X}_{ras}$ are predictors specified at the sub-area level, $\hat{X}_{ra}$ are predictors specified at the area level, and $\hat{X}_r$ are dummies specified at the regional (i.e. representative) level.

The MERF sub-area model can be written as follows:

$\hat{P}_{ras} = f\left(\hat{X}_{ras}.\hat{X}_{ra}, X_r\right) + \eta_{ra} + \varepsilon_{ras}$,

where $f(\ \ )$ is a random forest prediction function, $\eta_{ra}$ is an area specific random effect conditioned on the survey data $\hat{P}_{ras}$, and $\varepsilon_{ras}$ is a stochastic error term.

These two machine methods differ from traditional EBP methods and each other in three ways. The first is the use of random effects, which are present in the MERF model but not in standard XGBoost models. The random effects use an empirical Bayesian framework that essentially "updates" the survey estimates with the synthetic prediction using the auxiliary data. This appropriately gives the survey data more weight, relative to the predictions, in target areas where there are more surveyed households. The second is the calculation of uncertainty. Estimating uncertainty properly for gradient boosting models is still an area of active research, although the same type of cluster bootstrap approach used to estimate uncertainty for MERFs has shown promising results in simulations (Krennamir and Schmid, 2022, Merfeld and Newhouse, 2023). The third dimension on which these approaches differ is that the machine learning approaches use more flexible tree-based modeling approaches. The latter tend to be more accurate when there is sufficient training data because they are more flexible with regards to modeling interactions and are more robust to outliers. In general, the relative accuracy of these different methods depends on the sample and population data; the relationship between the nature of the sample and auxiliary data and the accuracy of estimates produced by different methods is not fully understood and is an active area of research.

Accuracy, however, is not necessarily the only criterion for selecting a method, as parsimony and explainability are also relevant in many settings (Efron, 2020). Parsimony refers to models that have a relatively small number of parameters that can be communicated, while explainability refers to the ease of communicating how the model works and understanding how changes in model predictions are influenced by ceteris paribus changes in individual predictors. Empirical Best Predictor (EBP) models are parsimonious, more explainable, and far simpler to communicate than machine learning models, since they are based on linear regressions coefficients. Nonetheless, linear regressions are not always easy to interpret. For example, it is common to include many variations of a single variable as predictors – current, annual mean, annual maximum, and annual minimum NDVI, for example. Interpreting the coefficients of current NDVI when controlling for these annual values is not straightforward, and even less so when also including other variables that are highly correlated with NDVI. In addition, the transformations common in linear mixed models make the magnitudes of the coefficients more challenging to interpret and communicate. The choice of whether to use machine learning or tree-based methods therefore depends on the data context, as well as the preferences of the intended users of these estimates.

### B. Selecting predictors and evaluating assumptions

As is probably clear, thousands of geospatial indicators can be created using publicly available data alone. For example, just for NDVI, one can calculate the average annual minimum, maximum, mean, and standard deviation; the current annual minimum, maximum, mean, and standard deviation; the current month's value; the previous month's value; and the same statistics at higher levels of aggregation (for example we can calculate these values at both the subarea and the area level). With such a wide array of possible predictors, a key potential problem emerges: overfitting.

While adding more variables will make the estimation "better" in the sample, it can lead to poor predictions out of sample by essentially modelling noise in the sample. Another way to think about this is, with enough covariates, one could model outcomes nearly perfectly in the sample even when there is lots of measurement error; the out of sample predictions, however, would likely be very inaccurate.

To help avoid this, many practitioners first use model selection procedures like the least absolute shrinkage and selection operator, commonly called lasso (or LASSO), especially when working with linear models (Tibshirani, 1996).[30] Lasso works by selecting a set of features (covariates) that offer the best predictions. Since LASSO has not yet been integrated into small area estimation packages, we recommend using LASSO to select the model, and then estimating the model with the selected covariates. To determine the number of variables to select, LASSO often relies on a process known as cross validation, where a model is fit on a subset of available training data and then predictions are generated on the rest of the training data, allowing for an estimate of accuracy. This process is repeated multiple times using different subsets and held-out subsets and the set of covariates that best predict, on average, across these different subsets are then used in the subsequent SAE model. Cross-validated LASSO has been implemented both in Stata and in the R glmnet package. However, Stata has implemented an alternative procedure which minimizes the Bayesian Information Criterion (Schwarz, 1978), based on coefficients that are not shrunk towards zero (Zhang et al, 2010). This variant of LASSO may therefore be better suited for the type of variable selection exercise considered in this context, in which LASSO is used purely for selecting predictors to use in a mixed model. Yet another variant of LASSO, implemented in the glmmLasso

[30] Many machine learning algorithms apply similar methods to help avoid overfitting. Many of these methods involve introducing randomness into the algorithm, though there are algorithms that incorporate lasso-like procedures, as well.

package in R, can include random area-specific intercepts conditioned on the data. Theoretically, this is an appealing procedure because of its consistency with the structure of EBP models. However, we know of no evidence that selecting models using mixed model LASSO procedure makes a large difference in practical settings. Regardless of the exact variant of LASSO used, LASSO avoids overfitting the sample data by selecting models with similar values of in and out of sample $R^2$. Both Stata and R allow one to specify variables that are not penalized when performing LASSO, which is useful for example to include the full set of regional dummies. In other words, the user can specify which variables will be automatically included in the final model.

The role of multicollinearity in model selection can be confusing. Multicollinearity occurs when two variables that are highly correlated are included in the model, such as average night-time lights and its square. This makes the estimated coefficients on these variables imprecise, because their conditional correlations are difficult to distinguish. However, multicollinearity does not increase the variance of the prediction itself.[31] This is because the accuracy of the prediction depends only on the full set of predictors used in the model, and not on the precision of the estimated coefficients for individual predictors. Dropping a collinear variable reduces the geometric space spanned by the predictors, which will slightly reduce in-sample predictive performance. If one is using a model selected by LASSO, with similar values for in and out of sample $R^2$, dropping collinear variables will lead to a slightly underfit model, which will slightly reduce the accuracy of the estimates. Removing collinear variables, however, can improve predictive performance if the joint relationship between the collinear variables is different in the sample used to estimate the model

[31] Mechanically, this occurs because of the estimated covariances between the coefficients of highly collinear predictors are large and negative.

than the rest of the population. Therefore whether to include or exclude highly collinear variables following LASSO depends on the context and judgment of the modeler.

The use of LASSO or other automated selection procedures is useful when using variants of linear models, such as the Empirical Best Predictor (EBP) model. Some newer SAE methods that use machine learning methods – like XGBoost as utilized by Chi et al (2022) and Merfeld and Newhouse (2023) – do not require selecting variables before estimating the model, because they have their own procedures to prevent overfitting. We provide examples of this below.[32]

### C. Estimating Point Estimates and Uncertainty

It is important to estimate a measure of uncertainty around the point estimate for a given outcome of interest, in part to verify that estimates are sufficiently precise to publish. Consider a one-fold empirical best predictor (EBP) model, estimated at the household level:

$$f(Y_{asi}) = \beta X_{as} + \gamma X_a + \eta_a + \varepsilon_{asi}, \quad (1)$$

where $f(Y_{asi})$ is a (possible) transformation of the outcome of interest $Y_{asi}$, for household $i$ in subarea $s$ in area $a$; $X_{as}$ is a set of subarea-level covariates; $X_a$ is a set of area-level covariates; $\eta_a$ is a conditional random effect, specified at the area level; and $\varepsilon_{asi}$ is a household-level error term. In practice, we recommend to also include higher-level dummies at representative areas; in Mexico, for example, one can include state-level dummies as predictors, while in Malawi, one could include district-level dummies as predictors.

[32] A method closely related to lasso is ridge regression. Unlike lasso, ridge does not select different features. Instead, it shrinks the size of the coefficients to prevent overfitting. However, we are not aware of any packages that allow for implementation of ridge regression in the SAE context. Another option, called elastic nets, is a combination of both ridge and lasso.

To generate point estimates of different indicators, software packages typically use Monte-Carlo approaches, where $\eta_a$ and $\varepsilon_{asi}$ are drawn L times from their estimated distributions. Outcomes, such as poverty rates, are calculated for each replication and then aggregated to target areas and averaged across replications. The advantage of this approach is that it can be used to obtain any indicator derived from the underlying population, including inequality measures. The default value of L in the povmap package is 50, although using L=100 is also common. An attractive alternative is to obtain analytic solutions of the expected value of outcomes without Monte-Carlo simulations, which greatly reduces computational costs and should, in theory, slightly improve the accuracy of the estimates. This has been implemented for estimating means and headcount poverty rates in the povmap software package, as discussed further below.

The key to calculating uncertainty is in the two random variables in equation (1): $\eta_a$ and $\varepsilon_{asi}$. In theory, these two variables could follow any distribution. In practice, however, they are typically assumed to be normally distributed.[33] How are these used to calculate uncertainty? By estimating the variance of each of these error terms and then then implementing a parametric bootstrap. Essentially, the above model is fit to the data, the variance of $\eta_a$ and $\varepsilon_{asi}$ are estimated, and then the two error terms are randomly sampled from these distributions to generate a new dependent variable $Y_{rasi}$ for all population households. These are then used to generate new indicators of "truth" for a particular replication. Similarly, the model is used to generate a new version of the outcome in the sample data, which is then combined with the geospatial indicator to generate new predictions. These predictions are then compared with the "truth" derived from the new dependent variable $Y_{rasi}$ assigned to all population households. After repeating this process many times, the

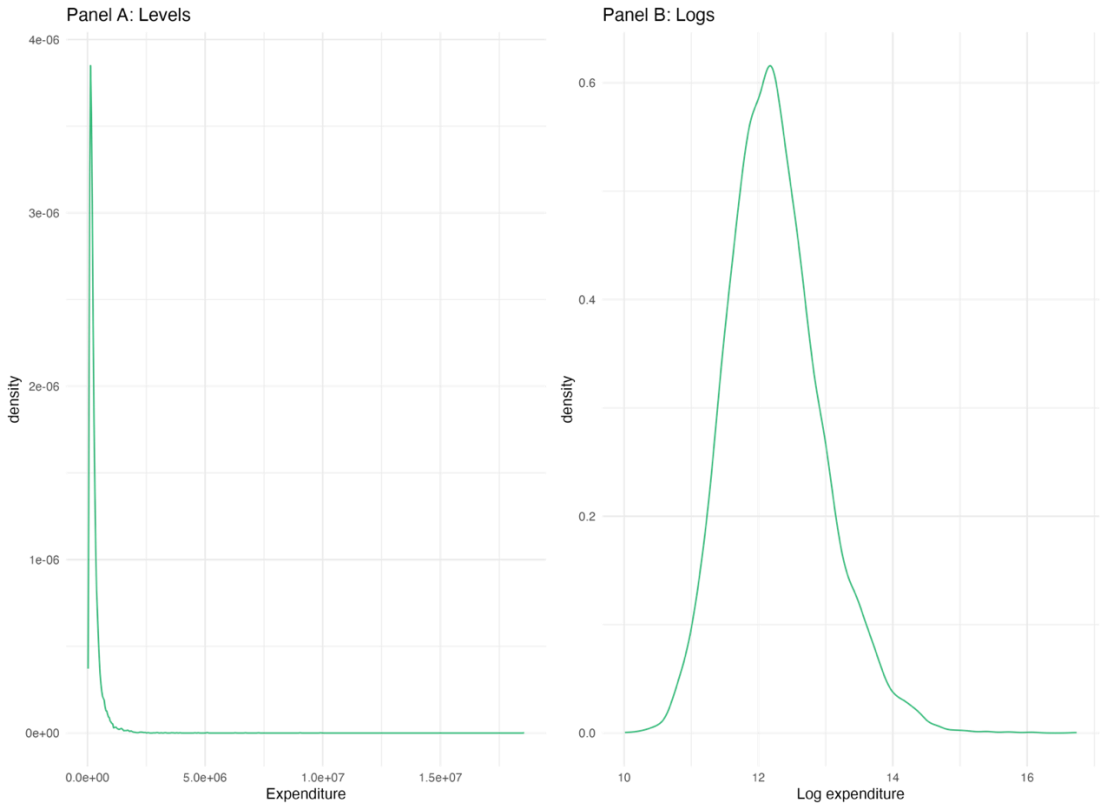
[33] Diallo and Rao (2018) study the possibility of skew-normal error terms. However, as of this writing, we are not aware of any software packages that allow for non-normal distributions.

procedure takes the average of the squared difference between the estimates and the "truth" derived from the full population across all replications, which is used as an estimate of the Mean Squared Error associated with the model. Common practice is to assume that the estimates are unbiased, which implies that the standard error is equal to the square root of the Mean Squared Error. The estimated standard errors can be multiplied by 1.96 to generate confidence intervals, and the Coefficient of Variation for each target area can be estimated as the ratio of the estimated standard error to the point estimate.

### D. Transformations

The EBP relies on the assumption that the area effect and idiosyncratic error term are normally distributed, both to generate point estimates and estimates of uncertainty. It is therefore common to transform the outcome variable to make these assumptions more palatable. Elbers et al. (2003) estimated poverty metrics by modeling welfare, defined as household-size adjusted income or consumption. As such, they recommended a log transformation, to make the welfare measure more normal. Figure 5 shows the results of such a transformation, using the 2019 Malawi IHS5. The left panel presents real per capita expenditures in levels. As is common with expenditure and income data, the distribution is highly right skewed, with a mass towards the left-hand side of the distribution; this is clearly far from normally distributed. The right panel, on the other hand, presents the distribution of the log-transformed real per capita expenditures. Although not perfectly normal, the resulting distribution is clearly much closer to normal than in the original levels. To put some numbers on the differences, we can compare the skewness and kurtosis of the distribution to those that apply to normally distributed variables (zero and three, respectively). In levels, the skewness is 17.3 and the kurtosis is more than 650. For the transformed variable, on the other hand, the skewness is 0.6 and the kurtosis is 3.8.

Figure 5: Expenditures in the Malawi IHS5




The two figures show the distribution of real per capita expenditures in the 2019 Malawi IHS5. Panel A presents the distribution in levels, while Panel B presents the distribution of log-transformed expenditures. The density estimates are unweighted.

While log transformations can work well for expenditure, a variety of transformations are implemented in the EMDI and povmap packages. This is useful for example for proportions, such as aggregating headcount poverty to the subarea level in the context of a subarea-level model. A common transformation for the case of proportions is an arcsine square-root transformation, which is sometimes referred to as the arcsin transformation or angular transformation. It is defined as:

$$f(Y) = sin^{-1}\left(\sqrt{Y}\right). \quad (2)$$

This type of transformation is appropriate for variables that range between zero and one, like headcount poverty.

There are other possible transformations, which we cover just briefly. The *povmap* package (Edochie et al, 2022) and *emdi* package (Kreutzmann et al., 2019) in R, for example, implement a Box-Cox transformation as the default, but also offer a log-shift transformation. The log shift adds a constant before taking the log. The value of the constant is determined analytically through a restricted maximum likelihood procedure, as described in Rojas-Perilla et al. (2020). Another option is a rank order transformation (Peterson and Cavanaugh, 2019), which forces a normal distribution for the outcome variable. Although this does not guarantee that the residuals are distributed normally, in practice this seems to work well for poverty (Masaki et al., 2020; Newhouse et al., 2022).

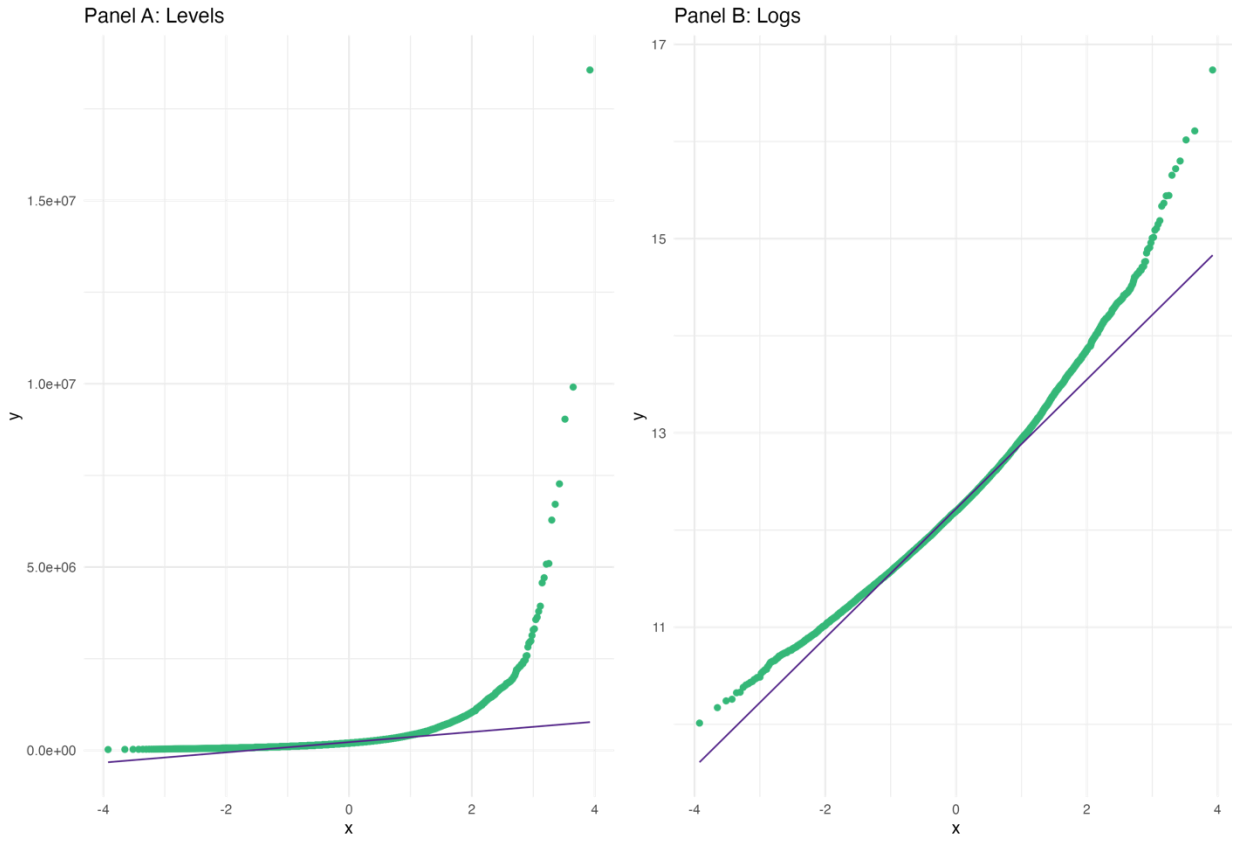
While all of these transformations can also be back-transformed, it is not always necessary to use transformations when solely estimating poverty headcount or stunting rates. Any monotonic transformation can be applied to the welfare distribution and threshold, negating the need to make the comparison in the original units. This approach also has the advantage, as mentioned above, of enabling simple analytic calculations instead of repeated Monte-Carlo simulations to generate point estimates of headcount poverty, which greatly reduces computational costs and slightly improves the accuracy of the estimates.[34]

It is important to assess whether the error terms appears to follow the assumed distribution – again, usually a normal distribution – by examining sample residuals. Kurtosis or skewness are common measures for this, as mentioned above. However, a useful visualization method is a QQ (quantile-

[34] Although repeated parametric bootstrap replications are still required to estimate uncertainty even if Monte-Carlo simulations are not required to generate point estimates.

quantile) plot. Since the quantiles of a true normal distribution are known exactly, it is possible to plot these theoretical quantiles of a normal distribution against the actual quantiles observed in the data. These plots are easily generated following model estimation using the qqnorm function in the R povmap or EMDI packages. Figure 6 presents QQ plots for expenditures in Malawi, using the exact same transformations as in Figure 5. If a variable were perfectly normally distributed, the scatter plot would exactly follow the purple lines in the two figures. The QQ plot for levels (Panel A) indicates much larger deviations from normality than the QQ plot for logs (Panel B) does. This is corroborated by the skewness and kurtosis, mentioned above.

Figure 6: Quantile-quantile plots of expenditure in the Malawi I5




Both figures are quantile-quantile plots for expenditures in the MalaIIHS5. The left figure presents expenditures in levels, while the right figure presents expenditures in logs.

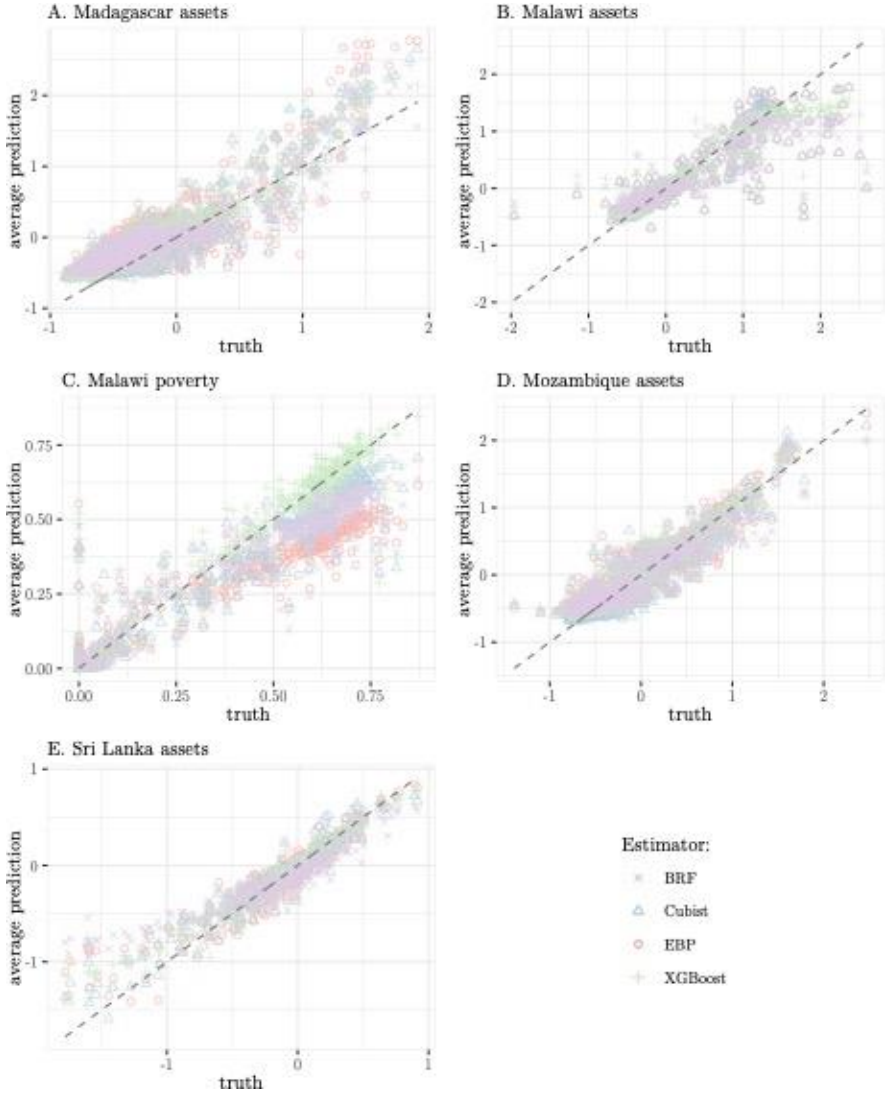
## E. Validation methods

Once we have our estimates, how do we know how accurate they are? To estimate accuracy, we have to compare the estimates to a benchmark measure of "truth." There are two general strategies used to measure accuracy, one less common method and one more common method. The less common method is census validation. If we have a census available, we can compare our estimates to the "truth" derived from the census. The definition of "truth" depends on the indicator. Many non-poverty indicators such as assets are directly measured in the census. Income and consumption are generally not, although there are a few exceptions. In these cases, one can impute a measure of welfare into the census and use this predicted welfare as "truth," with the understanding that measured welfare is noisier. One can add noise to this indicator to simulate the original noisier indicator. But care is required when evaluating different methods against a model-based benchmark, since the choice of modeling method used to generate the benchmark may affect the results. For example, using a benchmark measure of "truth" predicted using a linear regression model may overstate the performance of linear regression vis-à-vis tree-based machine learning.

Of course, a census is not usually available, which is what motivates the use of geospatial data for small area estimation methods in the first place! Nonetheless, census validation is the preferred approach to judge how well methods – particularly newer methods – perform. Consider, for example, Merfeld and Newhouse (2023). In that paper, the authors use census data from four separate countries – Madagascar, Malawi, Mozambique, and Sri Lanka – to evaluate the accuracy of an SAE procedure to predict a wealth index based on machine learning methods (XGBoost).

Figure 7 reproduces one of the figures in that paper. With knowledge of "truth" from the census, it is possible to plot the true value on the x-axis against the predicted values on the y-axis.

Figure 7: Census validation in Merfeld and Newhouse (2023)




The figure is a reproduction of Figure A3 in Merfeld and Newhouse (2023). Each figure plots the truth derived from a census on the x-axis and the estimates from SAE methods on the y-axis.

In addition to graphs, census data can also calculate statistics related to accuracy and precision. What are some of these statistics commonly used to measure accuracy in SAE? Correlations, both Pearson and Spearman, are common statistics. These are useful because Pearson and Spearman

correlations are usually closely related, and the latter is a useful measure for ranking the areas for targeting purposes. However, these are seldom used alone because it is theoretically possible to have a correlation of one even if all the estimates are wrong.[35] Other common statistics instead measure how far from truth each estimate is. Mean absolute deviation – defined as $\frac{1}{n}\sum_{i=1}^{n}|y-\hat{y}|$ – is one option, as is mean squared deviation – defined as $\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2$.

Accuracy is not the only important statistic, however, as it is also important to report measures related to precision. A common measure of precision is the estimated mean squared error (MSE), which is generated using the parametric bootstrap procedure described above when implementing EBP models. The standard error (SE), approximated as the square root of the estimated mean squared error, is also a useful indicator of precision. The width of the confidence interval, which is a simple product of the estimated standard error, conveys the same information as the SE, but in an arguably easier-to-digest format. Finally, it is useful to report statistics relating to the estimated coefficients of variation for the target areas, which is defined as the estimated standard error divided by the point estimate. For example, an area with an estimated poverty rate of 10 percent and a standard error of 5 percentage points would have a coefficient of variation of 0.5. The coefficient of variation, unlike the MSE or SE, has the important practical advantage of being independent of the scale of the dependent variable. For this reason, many NSOs use CV-based thresholds to determine if statistics should be suppressed due to lack of precision. However, we caution that the CV can be misleading when estimated outcomes are very close to zero or one.[36]

[35] As a simple example, suppose that $y = x - 1$. $x$ and $y$ will have a correlation of one, even though $x$ is never equal to $y$.

[36] An additional issue with using CV arises when modelling binary outcomes, like poverty. A poverty rate of 0.1 and 0.9 are mathematically similar, since we could instead model "non poverty." Yet, the CV will be much larger with the estimated average of 0.1 than 0.9.

Beyond accuracy and precision, it is also useful to measure the coverage rate relative to a benchmark measure of truth. The coverage rate indicates the share of areas for which the true value falls within the estimated confidence interval. It is therefore a way to assess how accurate measures of uncertainty are. We often calculate 95-percent confidence intervals, which means the coverage rate should be equal to approximately 0.95 if uncertainty is accurately estimated.

Most practical applications that use geospatial data for SAE are undertaken specifically because there is no recent census available. This means that, in practice, one cannot rely on census validation methods for these applications. This leads to the second main way to validate accuracy, which is cross validation. We have already discussed cross validation in the context of lasso, but here we provide a bit more detail.

There are many ways to implement cross validation, but we will use a single example: 10-fold cross validation. The idea behind cross validation is to insulate part of the data from the entire modelling process. In 10-fold cross validation, the data is randomly partitioned into 10 separate subsets, or folds. The entire workflow – from feature selection to estimation – is performed on just nine of these folds and then the resulting model is used to predict outcomes into the $10^{th}$ fold (the holdout sample). This same procedure is performed nine additional times, holding out each fold once. Predictions in the holdout sample from each iteration can then be compared to the actual value in each holdout sample, creating a measure of accuracy. There are multiple variations of this type of cross validation, including different numbers of folds. Another common method is "leave-one-out" cross validation, where just a single unit (or cluster) is in the holdout sample for each

iteration. Regardless of the type of cross validation, it is of vital importance that the holdout sample be completely insulated from the other folds in order for cross validation to work properly. Essentially this means that the holdout fold cannot be used for any type of estimation, including things like estimating means; it is only used to test the accuracy of out-of-sample predictions.

Cross-validation raises several thorny issues, and its behavior is complex and not fully understood. In general, cross-validation estimates an average prediction error, averaged across hypothetical datasets drawn from the underlying population (Bates et al., 2023). Yet all of these hypothetical datasets may differ in important ways from the full census population. For example, enumeration areas may be more geographically dispersed in many two stage samples than the population they are drawn from, if National Statistics Offices are averse to sampling nearby enumeration areas.

A second related issue involves the proper use of weights. Cross-validation in practice often does not account for sample weights, thus treating the unweighted sample as the population of interest. Samples, however, are often constructed by sampling enumeration areas with probability proportional to size, and therefore overrepresent larger areas. Samples may also over or under sample particular areas for reasons other than population. Failing to account for weights in the cross-validation procedure may also affect the accuracy of the reported results, though this behavior is not well understood in the cross-validation context.

A third issue with sample cross-validation is noise in the reference value due to sampling error, for example of households from within an enumeration area. The reference measure is therefore subject to sampling error, which in turn reduces estimates of model accuracy for reasons unrelated

to the accuracy of the model predictions when evaluated against the full population. In other words, sampling error can reduce estimated accuracy obtained from cross validation, even if the model predicted outcomes perfectly if all households were in the sample.

A fourth challenging issue is whether to use block cross validation, in which all households within some geographic area are assigned to a fold, instead of assigning individual households to folds. There is currently a methodological debate in the literature. Some studies argue that block cross validation is necessary to account for spatial autocorrelation in the data, while others argue that spatial cross-validation has no theoretical underpinning and should not be used for assessing accuracy (Wadoux et al., 2022).

One issue with block cross-validation at the target area level is that all predictions are, by construction, evaluated in areas outside the sample. This is different from many practical applications, in which sample data is drawn from within target areas. When sample data exist from within target areas, Bayesian and Empirical Bayesian methods can utilize the sample data as a prior that is updated by predictions, but unfortunately the benefit of this cannot be tested when conducting block cross-validation at the target area level. Aggregating block cross-validation at the EA and household level to higher levels is also problematic, because of leakage. Namely, the same sample data in that case is used both for model estimation and evaluation, after aggregating up estimates from different folds to higher geographic levels. Maintaining separate test and training samples within each target area avoids this leakage, at the cost of substantially reducing the sample for training, testing, or both.

Overall, cross-validation is a useful second-best option for assessing the accuracy of synthetic predictions in cases where census data are unavailable, with appropriate caveats and understandings of its limitations. Block cross-validation can be useful for assessing the accuracy of predictions into areas outside the sample but will likely underestimate performance for sampled areas. Cross-validation procedures cannot accurately evaluate the accuracy of predictions from Bayesian or Empirical Bayesian models for sampled areas. Results from cross-validation of sub-area or household-level models should not be aggregated across folds to the area level to assess the accuracy of target area estimates based on these models. Finally, simple cross-validation is not an appropriate tool for estimating uncertainty, although a nested cross-validation scheme appears to work better for variance estimation in many settings (Bates et al, 2023). Overall, it is useful to be cognizant of these limitations of cross-validation, with an understanding that the literature on this topic is still evolving.

A final useful heuristic for accuracy involves comparing small area estimates with direct estimates at a more aggregate regional level higher than the target area level, for which direct survey estimates are sufficiently reliable to publish. A graph can be useful to show discrepancies. A formal statistical test, based on Brown et al (2001), can be applied:

$$Z = \frac{(\hat{Y}_{sae,r} - \hat{Y}_{direct,r})}{\left(\sqrt{\widehat{MSE}_{sae,r} + \widehat{Var}_{direct,r}}\right)},$$

where $\hat{Y}_{sae,r}$ are the outcomes estimated from the SAE model, aggregated to the region level, prior to any benchmarking procedure. $\widehat{MSE}_{sae,r}$ is the estimated mean squared error of the small area estimates for region r. This can be approximated as:

$$\widehat{MSE}_{sae,r} = \frac{\sum_{a=1}^{A_r} w_{ra} \widehat{MSE_{ra}}}{{A_r}^2},$$

where $\widehat{MSE}_{ra}$ is the estimated mean squared error of the small area estimates in area a in region r, $A_r$ is the number of areas in region r, and $w_a$ are population weights rescaled to sum to $A_r$. We recommend obtaining $\widehat{Var}_{direct,r}$ using the Horvitz-Thompson approximation, as implemented in the R povmap and SAE packages, as standard cluster-robust variance estimates can seriously underestimate variance by failing to account for the positive correlation in outcomes across EAs within target areas.

However, interpreting these comparisons between model-based and direct estimates at the regional level must be done with care, for three reasons. First, if the sample does not cover all target areas, the direct and small area estimates will be based on results from different sets of areas. To address this, we recommend aggregating small area estimates only over sampled areas, so as to engage in an "apples to apples" comparison across the same set of target areas for each region. Second, it is useful to ensure that the same weights are used when aggregating from target areas to regions; often direct estimates use sample weights and small area estimates use population estimates from a recent census or estimates derived from geospatial data such as those provided by WorldPop. Use of different aggregation weights, like aggregating over different areas, exacerbates the discrepancies between direct and modeled estimates at the regional level, for reasons completely unrelated to the accuracy of the target area estimates.

Even when making proper comparisons, discrepancies between direct and model-based estimates should not be fully attributed to model-based error, since both random and non-random sampling error can also influence direct estimates. While it is tempting to regard direct estimates at the regional level as a measure of "truth" because they are reported as official statistics, unfortunately

direct estimates may not always accurately represent the underlying population, if for example particular enumeration areas within the target areas are effectively excluded from the sample due to local conflict. Overall, the limitations of cross-validation and comparisons with direct estimates as evaluation techniques underscore the importance of rigorous evaluation of methodologies using census data.

## V. Skills and tools needed to incorporate geospatial data into small area estimation

In this last section, we discuss some of the key tools needed to incorporate geospatial data into small area estimation. We focus on the R programming language because it contains both software for processing geospatial data and user-friendly packages for conducting small area estimation. Packages for small area estimation also exist in Stata, while geospatial processing can also be carried out in Python. Before discussing the actual extraction of geospatial data, we first go over two key geospatial data types: shapefiles and rasters.

### a. Shapefiles

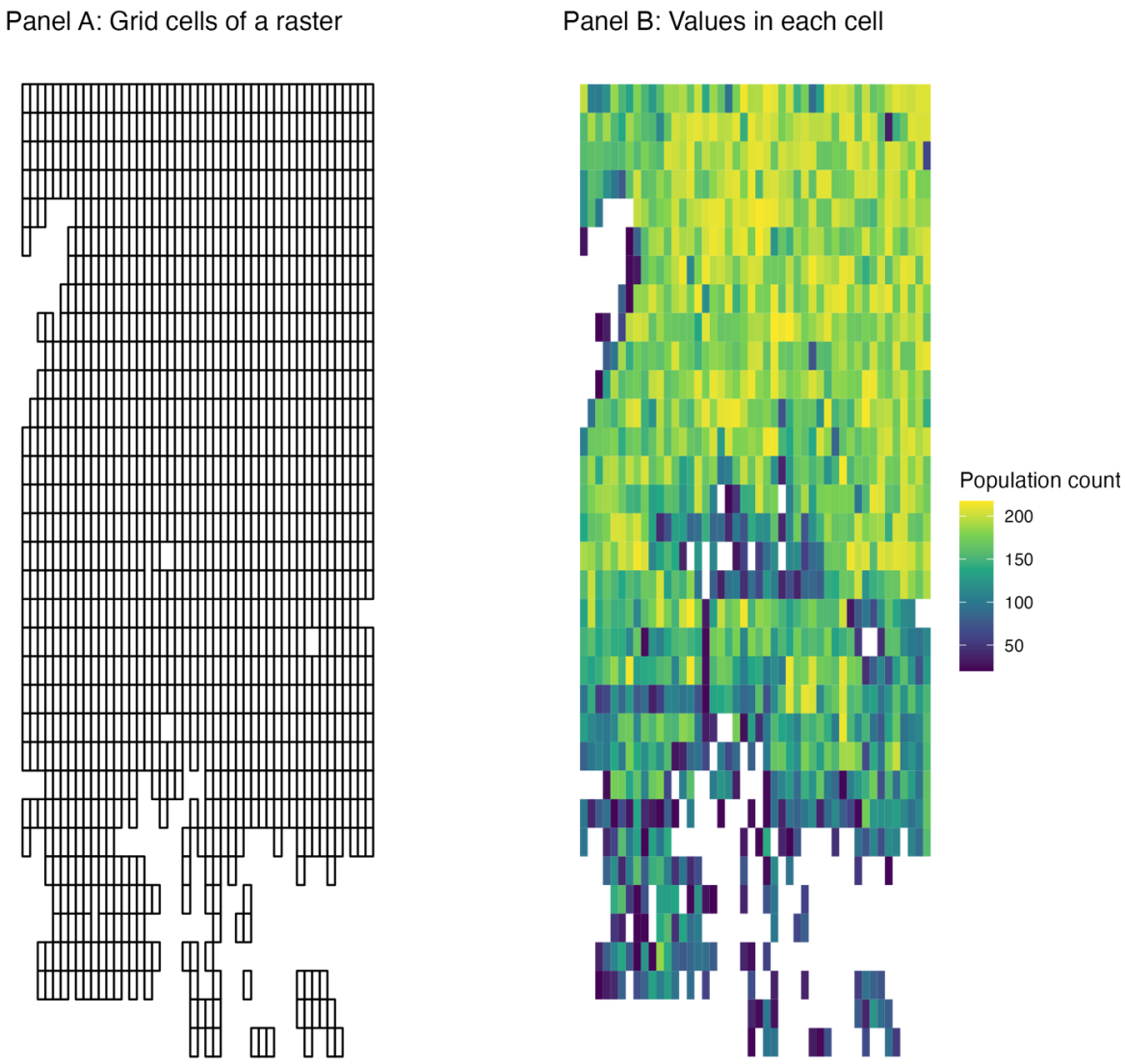
Shapefiles are one of the most common formats for geospatial data. These are familiar to almost everyone, even for users who have not used them directly; shapefiles make up many of the maps you see! Shapefiles are outlines of geographic areas or other features of interest – for example, a shapefile could be the outlines of buildings, although they are more commonly outlines of administrative areas. Shapefiles work by outlining these "polygons" with points. Consider a simple square; you can outline an entire square with just four points. A pentagon would require five points, a hexagon six points, and so on. Shapefiles are just a list of points – or vertices – that outline different features/polygons. It is not the geographic area of the feature that increases the size of shapefiles (in terms of computer memory), per se; instead, it is the number of vertices needed to

outline the shape. Some geographic boundaries are relatively smooth on the edges, which leads to fewer vertices, while others such as coastlines are complicated shapes requiring many more vertices. While shapefiles can include descriptive information about each feature – for example, some countries release shapefiles that include census-derived demographic information – it is the geographic nature of shapefiles that make them shapefiles.

### b. Rasters

Rasters are the second key data type to understand when working with geospatial data. Rasters differ geographically from shapefiles in one key aspect: shapefiles allow each feature to have different geographic shapes and sizes, while each cell of a raster has the exact same geographic dimensions. Panel A of **Error! Reference source not found.** is an example of raster grids. This comes from part of the largest city in Benin, Cotonou. Each grid cell is identical – in both outline and size – and these shapes cover the geographic area, with some gaps where there are no settlements (but note that the gaps are really made up of grid cells, they are just absent because they have no value assigned to them). Since each raster cell is of the exact same dimensions, the location of all grid cells can be inferred from knowing the exact location of one grid cell. For example, if we know the vertex (point) of the upper-left grid cell starts at the coordinates of (0, 0) and that each grid cell is 1-by-1, we can then immediately figure out the location of every grid cell in the entire raster.

Figure 8: Example of a raster




This is a raster from WorldPop's population estimates for Benin. The geographic area is part of Benin's largest city, Cotonou. The left figure shows just the raster grid cells, while the right figure also includes the value associated with each grid cell. The grid cells are squares; they have just been "squished" to allow a side-by-side presentation of the two rasters.

Since rasters are made up of identical grid cells, they are not generally used to display geographic shapes. Instead, they contain quantitative information of interest. Each grid cell could contain, for example, the value of precipitation, temperature, or NDVI. In the case of Figure 8, the raster comes from WorldPop and contains estimates of population counts. Each grid cell will contain a single value that corresponds to the variable it contains. Since this raster is of population, each grid cell

contains a single value that corresponds to the estimated population living in that grid cell. You can see this in Panel B.

In theory, the value associated with a raster can be anything. It is also common to have rasters that contain multiple variables – a raster "brick" or "stack" – that is the same idea as above. Another way to think of this is multiple rasters stacked on top of one another. The most common file format for raster files is GeoTIFF files, with an associated extension of .tif. But there are other formats that work similarly, like netcdf, which is used to store multiple variables. Thankfully, the R packages we discuss below can work with many of these different formats.

### c. Projections

One important detail to understand when working with shapefiles is how to deal with projections. The most common geographic format we are used to seeing in daily life is latitude and longitude. However, one issue with using latitude/longitude in geospatial work is that one degree of longitude corresponds to a different length (in kilometers) depending on the location of the globe. For example, at the equator, one degree of longitude (moving east or west) is around 111km. However, as you go north or south, this distance decreases until it is zero at the north and south poles. This means that latitude and longitude are not well-suited for visualizing maps if having consistent distances is important. This is where projections come in.

Projections are different ways of showing shapefiles or rasters on a flat surface and most projections are specific to a geographic area; this means there are thousands of different coordinate systems/projections. In our Benin example, the local projection is referred to as EPSG:32631 (WGS 84/UTM zone 31N). This is what is referred to as a coordinate reference system, or CRS.

One important detail about the UTM projections is that they are in meters, not degrees. This makes it more straightforward to calculate distances between points or areas of different shapefile features. The tools we discuss below make using projections a bit easier.

### d. R packages

Many of the tools associated with extracting geospatial data are intended to obtain zonal statistics from the raster that apply to shapefile polygons. We focus here on the R programming language, but there are other options, as well, including Python and its integration with specialized geospatial software like QGIS or ArcGIS.

The first package we recommend is the sf package (Pebesma, 2018).[37] The sf package makes reading shapefiles easy. More importantly, it makes it easy to work with shapefiles within R, as R will treat shapefiles like data frames. This allows the user to easily merge/join other data or perform mutations as with data matrices. In addition, plotting shapefiles with ggplot is straightforward. Consider the following code snippet:

```
library(sf)
library(tidyverse)
shape <- read_sf("LOCATION OF SHAPEFILE")
ggplot(data = shape) +
  geom_sf()
```

These three lines of code will load the shapefile into memory and plot the outline of the shapefile in R. It is then easy to export the map using ggsave.[38]

[37] https://r-spatial.github.io/sf/
[38] We load the `tidyverse` package in this example to allow us to use `ggplot`.

If the shapefile already contains a variable you want to map – like average temperature – you can slightly modify the above code to then color code that variable within the map. If the temperature variable is called *avgtemp*, then plotting is as simple as:

```
shape <- read_sf("LOCATION OF SHAPEFILE")
ggplot(data = shape) +
  geom_sf(aes(fill = avgtemp))
```

It is possible to work with shapefiles in other packages, but we have found the sf package to be indispensable when working with geospatial data in R.

The second package we use on a daily basis is the terra package (Hijmans et al., 2022).[39] We mainly use this to read rasters. In Figure 8, we displayed a raster of population values in Benin. Loading this into R is straightforward with terra:

```
population <- rast("LOCATION OF RASTER")
```

Plotting this is a bit more intensive due to how ggplot interacts with rasters, but extracting the data is more straightforward. For this, we use the exactextractr package (Baston, 2023).[40] While there are other packages that allow users to extract raster data to shapefiles, we have found that none compare to exactextractr in terms of speed, which can be very important for larger rasters or shapefiles.[41]

[39] https://rspatial.org/. We note that the raster package is also commonly used in geospatial work. The terra package is meant to replace raster, however.
[40] https://isciences.gitlab.io/exactextractr/
[41] The terra package also has an extract function. In our own experience, terra's extract function works much faster when trying to extract raster values to points, but exactextractr works faster when extracting to polygons. This could very well change in the future as the terra package continues to be updated.

The following code snippet will extract the values of the WorldPop raster to a shapefile for Benin. Note that we want the *total* of population for a given shapefile feature, so we will want to sum all of the values in the raster within a given feature.

```
shape <- read_sf("LOCATION OF SHAPEFILE")
population <- rast("LOCATION OF RASTER")
population <- exactextract(population, shape, fun = "sum",
                           append_cols = c("SHAPEID"))
```

The first two lines load the shapefile and the raster. The third line will extract the raster values to shapefile features by summing the raster population values within each feature.[42] Importantly, this third line will create a data frame, where each column represents one shapefile feature and each row represents an area the feature applies to. In order to be able to join this data with the shapefile, we add the append_cols option, which will add the shapefile feature ID to the resulting data frame. You will of course need to change the name of SHAPEID to the relevant identifier in your shapefile.

In Benin, we have an admin2 level shapefile. The unique identifier for each feature in this shapefile is admin2Pcod. In the previous code snippet, we would put "admin2Pcod" in place of "SHAPEID", and the following is an example of the first few rows of the resulting data frame in R:

| (row) | admin2Pcod | sum |
|---|---|---|
| 1 | BJ0101 | 310925 |
| 2 | BJ0102 | 140441 |

The sum column represents the estimated sum of population in the BJ0101 administrative area, according to Worldpop. If we were to do the same thing for many different geospatial variables, we would then be able to merge all these different variables together using the admin2Pcod unique identifier.

[42] It is also easy to substitute "mean" for "sum" to instead take the mean. Since we want total population, we use sum here. But in other contexts it may make sense to extract mean values of things like nightlight intensities.

There are currently packages available in R designed to streamline the acquisition of geospatial data. The Geodata package, for example, facilitates access to climate, crops, elevation, land use, soil, species occurrence, accessibility, administrative boundaries and other data. The Geolink R package, which remains under development, aims to automate the process of downloading raster data, extracting zonal statistics using exactextractR, and merging it with survey data. Packages and tools are evolving quickly and it is useful to periodically search the web to stay abreast of the latest developments.

### e. Conducting Small Area Estimation in R

In R, we prefer to use the ebp function of the povmap package (Edochie et al, 2023) to estimate linear SAE models. This package is a fork of the previously mentioned EMDI package with some new features to facilitate estimation of headcount and means. Povmap requires the user to specify "census" and "sample" dataframes, the model, the name of the variables indicating the target domains in the sample and population dataframe, as well as indicating several options. The ebp function estimates unit-level models using the empirical best predictor model pioneered by Battese, Harter, and Fuller (1988) and Molina and Rao (2010). This is an empirical Bayesian model that includes a conditional area random effect, which is conditioned on the sample data, as well as a standard idiosyncratic error specified at the unit level. Users can also specify their choice of transformation to be applied to the dependent variable, which as discussed above is important given the normality assumptions imposed on both the conditional random area effect and idiosyncratic error term in the model. In the case of log-shift and Box-Cox transformations, the relevant parameter will be determined through a restricted maximum likelihood procedure under the assumption that the transformed dependent variable is distributed normally.

The software also contains several options for benchmarking the estimates, so that the population-weighted averages match survey estimates at a more aggregate area that are considered to be representative. This can be convenient for ensuring for example that district-level estimates obtained through small area estimation are consistent with published state-level estimates obtained from survey data. More details are available in the three vignettes that document the EBP function in the EMDI and povmap packages. Finally, as mentioned above, a developmental version of povmap enables the analytic calculation of point estimates by setting L, the number of Monte-Carlo replications, to 0. This is currently only implemented for the calculation of headcount poverty and means when no transformation is specified and means when the arcsin square root transformation is specified. In these common cases, analytic calculations offer a large advantage in computational cost and speed over the standard Monte Carlo approach, as well as a slight improvement in the accuracy of the estimates. Povmap can also parallelize the estimation of the parametric bootstrap across multiple cores, which can greatly reduce computational time.

Povmap offers a number of diagnostic and reporting features that are convenient for users. It reports the marginal and conditional $R^2$, where the former only considers the predictor variables and the latter takes into account the conditional random effect. However, the between-area $R^2$, defined as $Corr\left(\bar{Y}_a, \hat{\bar{Y}}_a\right)^2$, or the squared correlation between predicted and measured average outcomes at the target area level in the sample, may be a more informative indicator of model accuracy. This is because the model predictions are ultimately aggregated to the area level to generate predictions at that level.

Povmap also reports the skewness and kurtosis of the residuals, and the qqnorm function will conveniently draw the quantile-quantile plot of the residuals, to allow for easy checking of normality. The write.excel function will output a few key diagnostics, the full set of estimates, and optionally model coefficients to an Excel spreadsheet. The documentation for povmap and EMDI detail these features.

Finally, povmap is under active development and has a number of experimental features implemented in a development branch. These include the ability to estimate two-fold models, the MSE_pop_weights option for estimating MSE properly for unit-context models, analytic point estimates to save computation costs, and more.

### f. Example code

As part of this paper, we have also included a fully reproducible example for Benin. You can download this repository on one of the author's GitHub page.[43] The readme file on the repository explains the general organization of the repository and also explains how to download the required data. The R scripts contain numerous comments to make them easier to understand.

## VI. Conclusion

Combining survey data with geospatial data has the potential to help fill in the spatial gaps from two stage sample surveys and obtain more precise and accurate estimates of important socioeconomic outcomes including SDG indicators. Publicly available geospatial indicators, which can be obtained from Google Earth Engine and other sources, are particularly effective in

[43] https://github.com/JoshMerfeld/saereplication

measuring urbanization, which is also highly correlated with a variety of socioeconomic indicators such as poverty.

Augmenting survey data with geospatial data can be done in many ways. Bayesian or Empirical Bayesian methods are attractive, particularly when using aggregate geospatial data, because they treat the sample data as a prior. This effectively combines the survey estimates and the model predictions, giving the sample more weight relative to the survey in target areas with larger sample sizes, and in applications when the model predicts relatively poorly. It also allows for the possibility that sampling error and model error at least partially cancel each other out, which becomes more likely when there are large errors in one or the other. In cases where they have been evaluated, Bayesian and Empirical Bayesian estimates tend to be generate similar predictions in practice, but the long tradition of Empirical Bayesian Predictor models for SAE and the existence of publicly available and well-documented software make the latter attractive. On the other hand, more advanced machine learning methods such as Extreme Gradient Boosting, while less transparent and more difficult to communicate, are also supported by publicly available and well-documented software and can generate more accurate predictions when the training data are sufficient to train a rich model.

Even within the set of Empirical Bayes models, there is an active research agenda on the pros and cons of different models. Our view is that the auxiliary data should be utilized at the lowest available data, which for geospatial data is typically the grid level or (hopefully) small administrative areas such as villages. These tend to give more accurate and robust estimates than area-level models, particularly when the selection of EAs may not be fully random, for example

due to conflict conditions that prevent some EAs from being surveyed. Area level models can also perform well but are a second best option in cases where lower-level geospatial data are not available. When applying area level models, it may be more important to mask the geospatial indicators so that they only apply to populated areas. Of the unit-level models, two feasible options are unit-context models and sub-area models. Although both use the same auxiliary data, unit-context model can perform slightly better (Newhouse et al, 2022). Perhaps this is due to modeling a continuous rather than a discrete variable.

Another area of active research is how best to use geospatial data to estimate inequality in small areas, which may require a different approach than is typically employed in small area estimation. This is because geospatial data can typically be linked to survey data only at a geographic level above the household, such as a grid or village, meaning that geospatial data cannot capture any variation in inequality within these grids or village. Modeling inequality in geospatial indicators at the grid or village level as a function of inequality in the geospatial data may be a feasible option, but we know of no research that has explore this question, and village level inequality measures may be quite noisy. This is another important question for future research.

The existence of Google Earth Engine, Planetary Computer, and similar platforms greatly facilitates the process of extracting zonal statistics and linking them to surveys. The R Geodata and Geolink packages, the latter of which is still under development, can further streamline this process when it is released. While the use of model-based estimates can introduce error, the evidence so far indicates that the benefit of adding predictions based on geospatial data to two stage sample surveys outweighs this error, making the model-based more accurate and precise than

direct survey estimates. Compared with direct estimates, augmented data often increases the precision of the survey by a factor of three to five. Obtaining a comparable increase in precision by expanding traditional data collection would cost millions of dollars. Our hope is that this guide helps data users and practitioners augment survey data with geospatial data, much of which is freely available, in order to improve measurement of socioeconomic outcomes.

References


Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, *603*(7903), 864-870.

Arima, S., Bell, W. R., Datta, G. S., Franco, C., & Liseo, B. (2017). Multivariate Fay–Herriot Bayesian estimation of small area means under functional measurement error. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(4), 1191-1209.

Ayush, K., Uzkent, B., Tanmay, K., Burke, M., Lobell, D., & Ermon, S. (2021, May). Efficient poverty mapping from high resolution remote sensing images. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 1, pp. 12-20).

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PloS one*, *12*(8), e0180698.

Baston, Daniel (2023). *exactextractr: Fast Extraction from Raster Datasets using Polygons*. https://isciences.gitlab.io/exactextractr/, https://github.com/isciences/exactextractr.

Bates, S., Hastie, T., & Tibshirani, R. (2023). Cross-validation: what does it estimate and how well does it do it?. *Journal of the American Statistical Association*, 1-12.

Battese, George E., Rachel M. Harter, and Wayne A. Fuller. "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association* 83, no. 401 (1988): 28–36.

Bell, W., Basel, W., Cruse, C., Dalzell, L., Maples, J., O'Hara, B., & Powers, D. (2007). Use of ACS data to produce SAIPE model-based estimates of poverty for counties. Census Report.

Bellow, M. E., & Lahiri, P. S. (2011). An empirical best linear unbiased prediction approach to small area estimation of crop parameters. *National Agricultural Statistics Service, University of Maryland*.

Brown, G., Chambers, R., Heady, P., & Heasman, D. (2001, October). Evaluation of small area estimation methods—an application to unemployment estimates from the UK LFS. In *Proceedings of statistics Canada symposium* (Vol. 2001, pp. 1-10). Statistics Canada.

Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535), eabe8628.

Casas-Cordero Valencia, C., Encina, J., & Lahiri, P. (2016). Poverty mapping for the Chilean comunas. *Analysis of poverty data by small area estimation*, 379-404.

Chao, S., Engstrom, R., Mann, M., & Bedada, A. (2021). Evaluating the Ability to Use Contextual Features Derived from Multi-Scale Satellite Imagery to Map Spatial Patterns of Urban Attributes and Population Distributions. *Remote Sensing*, 13(19), 3962.

Chi, Guanghua, Han Fang, Sourav Chatterjee, and Joshua E. Blumenstock. “Microestimates of Wealth for All Low-and Middle-Income Countries.” *Proceedings of the National Academy of Sciences* 119, no. 3 (2022): e2113658119.

Datta, G. S., Day, B., & Basawa, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, 75(2), 269-279.

Diallo, Mamadou S., and J. N. K. Rao. “Small Area Estimation of Complex Parameters under Unit-level Models with Skew-normal Errors.” *Scandinavian Journal of Statistics* 45, no. 4 (2018): 1092–1116.

Edochie, I., D. Newhouse, T. Schmid, and N. Wurz, “Povmap: Extension to the emdi package for small area estimation”, available at Comprehensive R Archive Network

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, *88*, S28-S59.

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350), 311-319.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, *71*(1), 355-364.

Engstrom, R., Newhouse, D., & Soundararajan, V. (2020). Estimating small-area population density in Sri Lanka using surveys and Geo-spatial data. *PloS one*, *15*(8), e0237063.

Engstrom, R., Hersh, J., & Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, *36*(2), 382-412.

Erciulescu, A. L., Cruze, N. B., & Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(1), 283-303.

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

Gething P, Tatem A, Bird T, Burgert-Brucker CR. 2015 Creating spatial interpolation surfaces with DHS data. *DHS Spatial Analysis Reports no. 11*. Rockville, MD: ICF International.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition. New Series*, *21*(4), 1-22.

Ghosh, M., & Lahiri, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 1153-1162.

Ghosh, M., & Lahiri, P. (1992). A hierarchical Bayes approach to small area estimation with auxiliary information. In *Bayesian Analysis in Statistics and Econometrics* (pp. 107-125). Springer New York.

Gualavisi, M., & Newhouse, D. L. (2022). Integrating Survey and Geospatial Data to Identify the Poor and Vulnerable: Evidence from Malawi, *World Bank Policy Research Working Paper 10257*.

Ha, N. S., Lahiri, P., & Parsons, V. (2014). Methods and results for small area estimation using smoking data from the 2008 National Health Interview Survey. *Statistics in Medicine*, 33(22), 3932-3945.

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.

Hawala, S., & Lahiri, P. (2018). Variance modeling for domains. *Statistics and Applications*, 16(1), 399-409.

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(2), 221-238.

Hersh, J., Engstrom, R., & Mann, M. (2021). Open data for algorithms: mapping poverty in Belize using open satellite derived features and machine learning. *Information Technology for Development*, 27(2), 263-292.

Hijmans, Robert J., Roger Bivand, Karl Forner, Jeroen Ooms, Edzer Pebesma, and Michael D. Sumner. "Package 'Terra.'" *Maintainer: Vienna, Austria*, 2022.

Hirose, M. Y., & Lahiri, P. (2017). A new model variance estimator for an area level small area model to solve multiple problems simultaneously. *arXiv preprint arXiv:1701.04176*.

Hobza, T., & Morales, D. (2013). Small area estimation under random regression coefficient models. *Journal of Statistical Computation and Simulation*, 83(11), 2160-2177.

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, and P. Lahiri. "Mixed Model Prediction and Small Area Estimfation." *Test* 15 (2006): 1–96.

Krennmair, P., and T. Schmid. "Flexible Domain Prediction Using Mixed Effects Random Forests." *Journal of the Royal Statistical Society Series C: Applied Statistics* 71, no. 5 (2022): 1865–94.

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*.

Lahiri, P., & Salvati, N. (2023). A nested error regression model with high-dimensional parameter for small area estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2), 212-239.

Lange, S., Pape, U. J., & Pütz, P. (2022). Small area estimation of poverty under structural change. *Review of Income and Wealth*, 68, S264-S281.

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

Lobell, David B., George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray. "Eyes in the Sky, Boots on the Ground: Assessing Satellite-and Ground-based Approaches to Crop Yield Measurement and Analysis." *American Journal of Agricultural Economics* 102, no. 1 (2020): 202–19.

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., & Ermon, S. (2023). Geollm: Extracting geospatial knowledge from large language models. arXiv preprint arXiv:2310.06213.

Masaki, T, D. Newhouse, A. \Silwal, A. Bedada, and R. Engstrom. "Small Area Estimation of Non-Monetary Poverty with Geospatial Data." *Statistical Journal of the IAOS*, no. Preprint (2020): 1–17.

McBride, L., Barrett, C. B., Browne, C., Hu, L., Liu, Y., Matteson, D. S., ... & Wen, J. (2022). Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning. *Applied Economic Perspectives and Policy*, 44(2), 879-892.

Merfeld, J., D. Newhouse, M. Weber, and P. Lahiri. "Combining Survey and Geospatial Data Can Significantly Improve Gender-Disaggregated Estimates of Labor Market Outcomes," *World Bank Policy Working Paper 10077*, 2022.

Merfeld, J., and D. Newhouse. "Improving Estimates of Mean Welfare and Uncertainty in Developing Countries." *World Bank Policy Working Paper 10348*, 2023.

Molina, Isabel, and J. N. K. Rao. "Small Area Estimation of Poverty Indicators." *Canadian Journal of Statistics* 38, no. 3 (2010): 369–85.

Morris, C., & Tang, R. (2011). Estimating Random Effects via Adjustment for Density Maximization. *Statistical Science*, 271-287.

Newhouse, David Locke, Joshua D. Merfeld, Anusha Pudugramam Ramakrishnan, Tom Swartz, and Partha Lahiri. “Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning.” *World Bank Policy Working Paper 10175*, 2022.

Newhouse, D., 2023. Small Area Estimation of Poverty and Wealth Using Geospatial Data: What have We Learned So Far?. *Calcutta Statistical Association Bulletin*

Otto, M. C., & Bell, W. R. (1995). Sampling error modelling of poverty and income statistics for states. In *American Statistical Association, Proceedings of the Section on Government Statistics* (pp. 160-165).

Pebesma, Edzer J. “Simple Features for R: Standardized Support for Spatial Vector Data.” *R J.* 10, no. 1 (2018): 439.

Pfeffermann, D., & Tiller, R. (2006). Small-area estimation with state: Space models subject to benchmark constraints. Journal of the American Statistical Association, 1387-1397.

Peterson, R.., and J. Cavanaugh. “Ordered Quantile Normalization: A Semiparametric Transformation Built for the Cross-Validation Era.” *Journal of Applied Statistics*, 2019.

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(1), 121-148.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., ... & Hsiang, S. (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, *12*(1), 4392.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., ... & Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127), 20160690.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

Tarozzi, A., & Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics*, *91*(4), 773-792.

Van der Weide, R, B. Blankespoor, C. Elbers, and P. Lanjouw. “How Accurate Is a Poverty Map Based on Remote Sensing Data?” *World Bank Policy Working Paper 10171*, 2022.

Vogt, M., Lahiri, P., & Münnich, R. (2023). Spatial prediction in small area estimation. *Statistics in Transition new series*, 24(3), 77-94.

Wadoux, A. M. C., Heuvelink, G. B., De Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692.

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., ... & Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, *115*(14), 3529-3537.

Yoshimori, M., & Lahiri, P. (2014). A new adjusted maximum likelihood method for the Fay–Herriot small area model. *Journal of Multivariate Analysis*, 124, 281-294.

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.

You, Y. "Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling." *Survey Methodology* 47.2 (2021): 361-371.

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.