On the use of geospatial data in small area estimation

Nikos Tzavidis[1,a], **Luciano Perfetti Villa**[1,b], **Vasilis Chasiotis**[2]

[1]Department of Social Statistics and Demography & Southampton Statistical Sciences Research Institute, University of Southampton, UK
[2]Department of Statistics, Athens University of Economics and Business, Greece
[1,a]n.tzavidis@soton.ac.uk, [1,b]L.Perfetti-Villa@soton.ac.uk, [2]chasiotisv@aueb.gr

Abstract

In an era where emphasis is placed on reducing operating costs whilst increasing the timeliness and granularity of survey estimates, it is natural to think how to make best use of alternative data sources beyond traditional sources of population data, e.g., from censuses. The use of administrative data has been the focus of extensive theoretical and applied research, mainly in data rich contexts where administrative data are available. Census and administrative data are, however, not available or regularly updated in many parts of the world. In this short paper, we discuss the use of geospatial data as a source of auxiliary information in model-based small area estimation. Using geospatial data in small area estimation is not new. For example, the seminal paper by Battese, Harter, and Fuller, 1988 used geospatial covariates in the nested error regression model. The growing availability, easy access, global coverage, frequent updates, and quality of remote sensing data has led to renewed interest in their use in model-based estimation. Their use however is not without challenges. How to process geospatial data ready for estimation, how to best integrate survey, census, and geospatial data, how to specify and build a small area model, and how to use geospatial data for intercensal updating are areas of current research interest. We study these topics by reviewing relevant literature and using evidence from three applications with access to data of varying detail. Specifically, the first application presents an ideal case with access to the households' geolocations and to census microdata from both a recent and an older census. In the second application, census microdata are not available, while in the third application, we have access only to area-level information. In all applications, geospatial data are used as supplementary or the only source of auxiliary data. We present a summary of current research findings with an emphasis on providing guidance to practitioners. Although the focus of the applications is on the estimation of general parameters of income and poverty indicators, the conclusions have broader applicability.

Keywords: Alternative data; data integration; intercensal updating; poverty mapping; spatial scales.

1 Introduction

Model-based and model-assisted small area estimation is commonly implemented with the aid of statistical models and auxiliary information from population data, for example censuses. Increasing demand for estimates of general parameters means that analysts must increasingly rely on population-level, e.g., census micro-data. Relying on census data can be restrictive. Censuses are less frequent in many parts of the world and are usually only conducted every ten years. Using outdated population data in the intercensal period relies on strong assumptions about the distribution of the census variables over this period.

The increasing availability, improved quality, frequency, and coverage of remote sensing data raises the question of the role that such data sources can play in small area estimation. Chi et al. (2022) have

already developed a methodology that relies on data from alternative sources. Advances in the processing, global coverage, frequency, and free access of remote sensing data have created renewed interest in their use as a source of auxiliary information in small area models, e.g., Van der Weide et al. (2022), Edochie, Newhouse, Würz, et al. (2024), Newhouse et al. (2025). This is in contrast to data, for example, from mobile networks that are not easily accessible. Using geospatial data in small area estimation is challenging. By definition, zonal statistics, i.e., summary statistics of remote sensing data, are computed at an aggregate spatial scale, e.g., a grid cell or an administrative unit. Therefore, zonal statistics are contextual predictors and act only as proxies of household characteristics. The predictive ability of geospatial data depends on the type of outcome we are interested in predicting and is application specific. The spatial scale to be used for the integration of geospatial and survey data, model specification, and model-building are all topics of current research interest.

Although geospatial data have been successfully used in several applications in several countries, we cannot assume that the same geospatial variables and model specifications will be equally predictive in other countries. The types of geospatial data and their utility in countries that are data-rich are also the focus of current research. In an era where emphasis is placed on reducing operating costs, making best use of alternative data sources becomes of paramount importance. Experiences in data-scarce settings offer valuable lessons in data rich settings. Data-rich settings also offer the opportunity to compare geospatial-based estimates against what are considered to be ”gold standard” estimates that use census data. Last but not least, a compelling reason for using geospatial data is that they offer a natural approach to updating the estimates in off-census years.

In this short paper, we present current research evidence on the use of geospatial data in small area estimation. The evidence we present comes from reviewing the literature, and focusing on three applications, part of our research, with access to data of varying detail. Specifically, the first application presents an ideal case with access to the geolocations of sampled households and to census microdata from both a recent and an older census. In the second application, census microdata are not available, while in the third application, we had access only to area-level data. The three applications allow us to explore research questions around the following themes: (a) integrating survey and census with geospatial data, (b) model specifications and model building, and (c) intercensal updating. Although we do not present detailed results, we summarise the key findings with emphasis on providing guidance for practitioners. The remainder of the paper is structured as follows. In Section 2 we introduce geospatial data and describe how to go from raster data to constructing covariates (predictors) and how to integrate geospatial data with survey and census data depending on the information that is available about the geolocations of the units. Section 3 focuses on three model specifications namely, unit-level, unit-context and area-level models, and on variable selection and data quality issues when using geospatial data. In Section 3 we also summarise the evidence from the three applications. Finally, Section 4 summarises the main findings and offers guidance for practitioners.

2 Geospatial data and small area estimation

The term geospatial data is collectively used to indicate information derived from satellite imagery, remote sensing instruments, and other spatial data collection systems that provide measurements at high spatial resolution with near-global coverage. The initial input for geospatial products is satellite imagery, which captures the Earth’s surface across multiple spectral bands, from visible light to infrared and microwave frequencies. These multispectral images undergo a series of processing steps to produce geospatial indicators of interest. First, the spectral bands are decomposed and calibrated

to correct for atmospheric interference and sensor characteristics. Classification/machine learning algorithms are then used to detect specific features of interest, such as building footprints, land cover types, or crop extent. The spectral information can also be transformed into smoothed surfaces, for example, by measuring the intensity of nighttime light emissions or by computing vegetation indices from the ratio of near-infrared to visible light reflectance. The resulting measurements are typically distributed as raster layers, where each pixel stores the value of the derived indicator at a fixed spatial resolution. The resolution varies considerably across geospatial products and determines the finest spatial scale at which meaningful variation can be observed. For example, Landsat imagery is captured at 30 metres per pixel, VIIRS nighttime lights at approximately 500 metres, and climate reanalysis products such as ERA5 at roughly 30 km. Repositories such as Google Earth Engine, Microsoft Planetary Computer, and WorldPop provide access to pre-processed geospatial layers with global coverage. Aggregated values of these layers using different summary functions and spatial scales (e.g., administrative areas) form the so-called geospatial zonal statistics which are used as auxiliary information (predictors) in statistical/machine learning models.

Geospatial data sources have become increasingly accessible over the past two decades, both through publicly funded programmes such as the Copernicus and Landsat missions and through the efforts of private organisations such as Google and Meta (Merfeld et al., 2023). Using geospatial data in small area estimation is not a new idea. The application in the seminal paper by Battese, Harter, and Fuller, 1988 used remote sensing variables as predictors in a unit-level small area model. However, the increased availability of and easy access to geospatial data has opened new opportunities for statisticians working, among other subfields, in model-based small area estimation (SAE), particularly in settings where traditional sources of auxiliary data, such as census and administrative microdata, may be outdated, incomplete, or not accessible. The lack of recent censuses in several parts of the world is a key reason for using geospatial auxiliary data and machine learning algorithms (random forests) to produce population estimates for high-resolution grids (e.g., 100 $\times$ 100 m or 1 $\times$ 1 km grids) (Stevens et al., 2017), (`https://www.worldpop.org/wp-content/uploads/2022/10/top-down-tutorial.html`). Geospatial data are also a key source of auxiliary information for Meta's global estimates of average wealth at 2.4 km^2 resolution (Chi et al., 2022). Although such products have been largely developed outside the core of the small area literature, their popularity shows that the use of geospatial data in applications of small area estimation has utility for practitioners. Commonly used geospatial variables in SAE applications include building counts and density (e.g., from Google Open Buildings or Microsoft Building Footprints), nighttime light intensity from the VIIRS instrument, land cover classifications from MODIS, elevation and slope from digital elevation models, climate variables such as temperature and precipitation, and distance-based features derived from OpenStreetMap and other geographic databases. The range of potentially useful geospatial products continues to expand as new satellite missions, data providers, and processing methods become available.

Unlike other alternative data sources, e.g., mobile phone data, several sources of geospatial data are freely available, frequently updated and have global coverage. However, careful use of geospatial data as predictors in models is needed. This is because, depending on the application, geospatial data do not always provide direct measurements about the unit of measurement (e.g., the household) but only about the context within which the unit of measurement is located. Hence, geospatial data act as proxies for household characteristics. We return to this, in our view, important point later in this paper.

2.1 From rasters to covariates: Integrating survey and geospatial data

A key step in using geospatial data for SAE is creating zonal statistics by aggregating the raster values at pixel-level to spatial units of interest. The aggregation involves computing summary statistics, most commonly the weighted average, for all pixels that fall within the spatial unit. The weights correspond to the fraction of each pixel covered by the corresponding polygon, which ensures that the partial coverage of the boundary pixels is properly accounted for (Baston, 2023). Because survey and census data are typically georeferenced at administrative area boundaries, it is common for raster values to be aggregated at that level. However, this does not necessarily have to be the case. How we decide to aggregate the raster values also depends on the information available about the geolocation of the units in the survey data. Figure 1 illustrates this point by showing how to integrate household data with geospatial data with survey data depending on the information available about geolocations in the survey data. The plot on the left (original raster) shows the case where the households' geolocations are available (red points). In this case, an alternative to administrative-level aggregation is to compute zonal statistics within a buffer zone around each household's coordinates (blue circles). For example, one may calculate the average intensity of nighttime lights or the average density of buildings within a radius of 1 km for each household. This approach attempts to produce zonal statistics that are close to, albeit still contextual, unit-level geospatial data. The choice of buffer radius introduces a similar trade-off to the choice of grid resolution: smaller buffers capture more localized conditions but may be noisier, while larger buffers smooth the spatial variation. The plot on the right (aggregated to Enumeration Area (EA)) shows a more common case where the exact geolocation of the households is not available (red points are illustrative of household locations in EAs). A possible solution in this case is obtaining zonal statistics at the lowest possible spatial level available (e.g., EAs). All households within the same EA will be assigned the same covariate values. This results in a smoother version of the nightlights variable on the right hand side plot. We return to the point regarding the information on geolocations that needs to be available in survey and population data later in this paper when we discuss different model specifications.

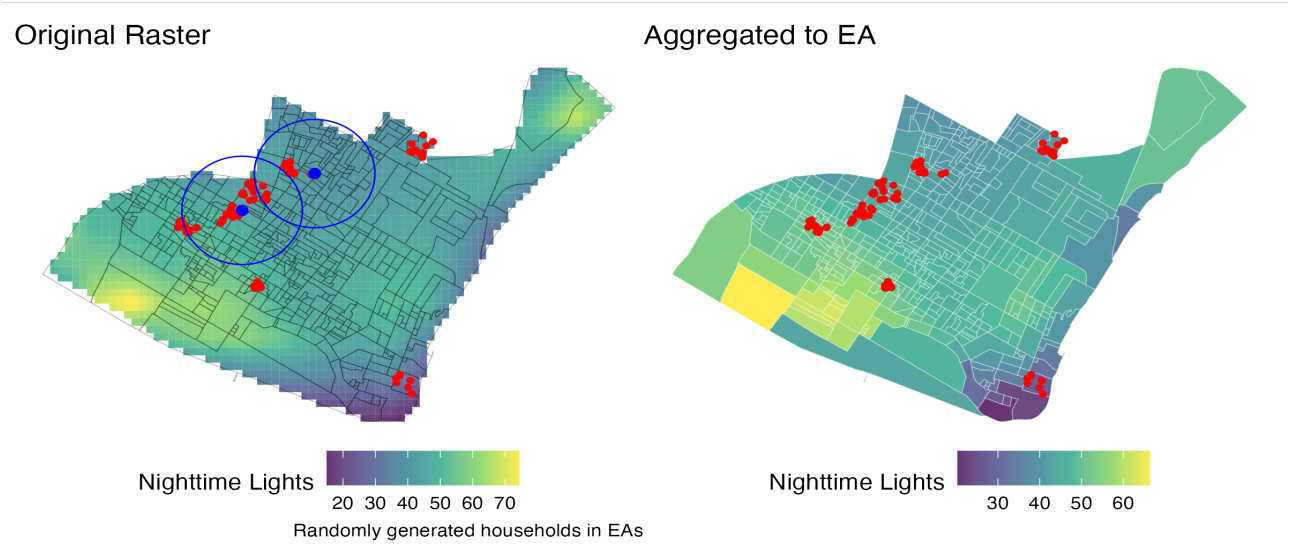

Figure 1: Nighttime Lights: Raster and aggregation to EA.

The choice of aggregation level for geospatial variables is a practical decision with methodological consequences because this will impact their subsequent use in models/algorithms for estimation/prediction purposes. We argue that if possible, analysts must be given full flexibility for producing zonal statistics. For example, analysts may decide to aggregate geospatial data to a variety of spatial supports, including regular grids of varying resolution (e.g., hexagonal or square grids at 1, 5, or 10 km^2) or administrative boundaries at different hierarchical levels (enumeration areas, sub-districts, or districts). This choice has consequences due to the Modifiable Areal Unit Problem (MAUP), which implies that the relationships between the geospatial predictors and the outcome may change depending on the spatial scale at which the predictors are measured (A. E. Gelfand, Zhu, and Carlin, 2001; Trevisani and A. Gelfand, 2013). Different geospatial variables may also exhibit different sensitivity to aggregation. For example, climate variables such as precipitation and temperature can vary considerably because the underlying spatial processes operate at different scales.

3 Model-based SAE with geospatial auxiliary information

The source of evidence presented in this section is current research on the use of geospatial data for small area estimation of income and poverty indicators. The estimation targets are general small area parameters including averages and poverty indicators (e.g., the mean consumption and the head count ratio). The model specifications below are presented in their simplest form, but, of course, several extensions to these specifications exist in the small area literature. We draw our evidence from three applications with access to data of varying detail. One application is from collaboration with the World Bank in Mozambique, the second application from collaboration with the Office for National Statistics in the UK, and the third application uses data from Greece, part of a wider collaboration between the World Bank and Eurostat on poverty mapping in Europe.

In the Mozambique application, we used data from the IOF 2022 household budget survey (15,382 households across 149 in-sample districts, with household geolocations available), the IOF 2019/20 survey (13,656 households, 154 districts), census microdata from the 2017 census (6,159,244 households with WorldPop 2022 population projections, 161 districts), 2007 census data (4.6 million households, 149 districts), and over 50 geospatial raster layers from WorldPop, Google Earth Engine, and Microsoft Planetary Computer. In the case of the UK application we used data from the Family Resources Survey (FRS), geospatial covariates aggregated at 100 $\times$ 100 m grids alongside administrative data such as benefits claimant counts, median house prices, and council tax bands, among others. In the application to data from Greece, we had access to area-level data from the EU-SILC survey for several years starting in 2010, geolocated information at the postcode level and access to corresponding geospatial data. The Mozambique application provided an ideal data scenario with census micro-data and household geolocations for processing raster data to various spatial scales. The availability of census data from 2007 allowed us to explore methods for intercensal updating. The UK data set-up was less favourable due to the lack of access to census microdata. However, unlike the case in Mozambique, administrative data were also available in this case. Finally, the application to data from Greece was the least favourable in terms of data access and exploring different model specifications.

3.1 The unit-level model

The unit-level nested error regression model, originally introduced by Battese, Harter, and Fuller (1988) and extended for estimating general parameters through an Empirical Best Predictor (EBP) by Molina and Rao (2010), is a widely used model specification for SAE. The ideal data scenario for

estimating general small area parameters assumes the availability of unit-level survey and population (e.g., census) microdata and the survey being conducted close to the census date. The predictors in the survey and census data are assumed to be measured using the same definitions. Under this data scenario, and assuming that the model has been built carefully, the unit-level model is a preferred method that has been shown to perform well. In practice, however, the ideal data scenario is rarely met because secondary analysts do not usually have access to census microdata and in the best case, censuses are only conducted every five to ten years.

The unit-level model specifies the relationship between a welfare variable y_{ij} (e.g., household income or consumption expenditure) for household j in area i and a vector of covariates $\boldsymbol{x}_{ij}$ as

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \tag{1}$$

where β is the vector of regression coefficients, and in the most common case, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ is an area-level random effect capturing unobserved heterogeneity across areas, and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$ is the household-level error. A transformation of the response variable, $h(y_{ij})$, such as the logarithmic-shift or data-driven Box-Cox transformation (Rojas-Perilla et al., 2019), is used when the distributions of the model residuals do not follow those assumed under the model, as is typically the case when modelling the raw income or consumption variable. Model (1) is estimated using the survey data. The target quantities in poverty mapping applications include the head count ratio (HCR), the poverty gap and inequality measures. Molina and Rao (2010) proposed the EBP which works by generating simulated/synthetic populations under the model, computing the target parameter in each simulated population, and averaging over the simulation replications. As mentioned above, the method assumes that the predictors $\boldsymbol{x}_{ij}$ are also available for each household in the population. Geospatial predictors can also be brought in the model as contextual variables alongside census microdata. Contextual variables can help explain part of the unobserved heterogeneity between areas and provide a valuable data source that we believe analysts should explore also under the unit-level model framework.

In the application to data from Mozambique, the EBP with the 2017 unit-level census data and 2019/20 unit-level survey data served as the gold standard for model-based estimation. At the district level, the EBP estimates of mean consumption and HCR under this model showed high Spearman correlations with direct estimates at both the district and province levels. Comparison at the province level is used because direct estimates at this spatial scale have acceptable statistical reliability. When the 2007 census was used instead, assuming that there was no census in 2017, the correlations between model-based and direct estimates declined. In this case, model-based estimates of average consumption in certain areas were underestimated, which also led to higher model-based estimates of HCRs than the corresponding reliable direct estimates. This pattern was particularly evident in urban areas. We suspect that this is caused because the support of the covariates available from the 2007 census is different from that from the 2017 census. Therefore, current evidence suggests that the use of outdated census data must be used with caution because they can affect the quality of small area estimates. In the applications with the data from the UK and Greece, we had no access to census microdata so producing EBP under the unit-level model estimates was not possible.

3.2 The unit-context model

The use of remote sensing data as auxiliary information in SAE dates back to the foundational work of Battese, Harter, and Fuller (1988), who used crop classifications at pixel-level from land observatory satellites, aggregated to segment level, as covariates to predict the average hectares by crop type

for segments in Iowa counties. The unit-context model (see also Nguyen (2012)) is a nested error regression model used in settings where the predictors are not measured at the unit level but are available instead at an aggregate spatial unit g (such as a grid cell, enumeration area, sub-district, etc.) within area i. In the simplest case, the unit-context model is specified as

$$y_{igj} = \boldsymbol{x}_{ig}^T\boldsymbol{\beta} + u_i + e_{igj}, \tag{2}$$

where $\boldsymbol{x}_{ig}$ denotes the vector of predictors aggregated to spatial unit g, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ is a random effect at the area-level and $e_{igj} \overset{iid}{\sim} N(0, \sigma_e^2)$ is the error term at the household-level. Note that under model (2) all units within the same spatial unit g share the same covariate values $\boldsymbol{x}_{ig}$, while the outcome variable y_{igj} remains measured at the unit level. This model specification is not unique to the use of geospatial covariates. Similar model specifications have been used when the source of auxiliary information is administrative data but no linkage between the survey records and the administrative records exists. This unit-context specification has been adopted in a number of applications using geospatial data for poverty mapping, including Masaki et al. (2022), Newhouse et al. (2025), and Edochie, Newhouse, Tzavidis, et al. (2025). Small area estimation of general parameters under this model also requires access to population sizes for each spatial unit g for the purposes of constructing the appropriate number of predictions to be used for deriving the small area estimates. If a recent census is available, and information about the geolocations of the households in the census is available too, these population sizes are readily available. However, exact geolocated information for all population units is rarely available. If access to such data is not possible also because there is no recent census, the population sizes can be approximated by using population projections from gridded population data such as the data produced by the WorldPop team at the University of Southampton. This situation is not unique to the use of geospatial data under the unit-context model. Deriving EBP estimates of general parameters under the unit level model also requires access to the population sizes for the target areas. However, getting access to the population sizes for higher level administrative units may be more straightforward than gaining access to population sizes for smaller spatial units.

Compared to the unit-level model, the use of the unit-context model has been criticized see e.g. Corral, Henderson, and Segovia (2025). Because all households within the same spatial unit share the same covariate values and in poverty mapping applications geospatial variables act only as proxies for household-level characteristics, it is reasonable to expect that the predictive ability of the unit-context model will not be as good as that of the unit-level model, possibly leading to higher unexplained variance and to less precise small area estimates. Despite this, the unit-context model can be a valuable tool for practitioners working in data limited settings. Its appeal lies in its applicability when census microdata are outdated and/or not accessible. As we outline below, our current view about the performance of the unit-context model is that it is application/context specific, and as a result, this model specification must be used with great caution.

In the case of the application in Mozambique, estimates under the unit-context model, with LASSO-selected geospatial covariates, achieved high Spearman correlation coefficients with direct estimates at the district and province levels for both small area means and HCRs. As expected, the predictive power of the unit-context model was lower than that of the unit level model and the corresponding MSEs generally larger than the MSEs of EBP estimates under the unit-level model. However, this was the case under the ideal data scenario for the unit-level EBP, i.e., when recent census data are available. When the estimates under the unit-context model were compared with the unit-level EBP estimates produced with the 2007 census data, the unit-context estimates were preferable. These

results indicate that in the case of Mozambique, carefully-selected geospatial covariates can provide reliable estimates at the district and province levels, and represent a clear improvement over the use of outdated census data. The aggregation level at which geospatial predictors are computed impacts on model performance. In Mozambique, geospatial predictors aggregated at coarser administrative levels (district, Posto) yielded estimates with lower RMSEs compared to the use of geospatial predictors at high-resolution hexagonal grids. This suggests that aggregation to finer grids introduces noise without improving predictive power. Using variable selection, e.g., via LASSO, is recommended when many geospatial layers are available, as it identifies which variables are important at each spatial resolution and addresses multicollinearity among correlated raster-derived covariates.

The above conclusions change when assessing the results from using the unit-context model in the application with the UK data. In this case, the unit-context model using geospatial covariates aggregated at 100 $\times$ 100 m grids performed poorly when used to estimate income-type measures for UK Local Authority Districts. In the UK, the rural/urban income difference was not explained well by geospatial variables. Our conjecture is that in developed countries the spatial variation in income is driven by socioeconomic factors (e.g., employment, benefits, housing costs) that are not captured well by remote sensing data alone. Including administrative data at low spatial scale improved the fit of the unit-context model, but the small area estimates remained unsatisfactory when compared to reliable direct estimates. In the application to data from Greece, we did not use the unit-context model because we worked only with area-level data.

3.3 The area-level model

The Fay-Herriot model (Fay and Herriot, 1979) is the standard area-level model in SAE. Unlike the unit-level and unit-context models which operate at the household level, the Fay-Herriot model is specified directly with area-level aggregates. The Fay-Herriot model is well suited for use with geospatial data because predictors, $\boldsymbol{x}_i$, are naturally linked to direct estimates at the area level. It comprises two parts. The first part is the sampling model which describes the relationship between the direct estimates $\hat{\theta}_i^{Dir}$ of the target area parameter with the true values θ_i,

$$\hat{\theta}_i^{Dir} = \theta_i + e_i, \quad e_i \overset{ind}{\sim} N(0, \sigma_{e_i}^2), \tag{3}$$

where e_i denotes the sampling error with variance $\sigma_{e_i}^2$, estimated under the sampling design. The second part is the linking model that relates the true area parameter to area-level covariates $\boldsymbol{x}_i$,

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i, \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2), \tag{4}$$

where u_i is the area-level random effect. Estimation of the Fay-Herriot model assumes that the variances of the direct estimates are known. The estimation of the sampling variances under the design is important because this determines the weight placed on the direct and the regression synthetic estimators used to construct the EBLUP.

Transformed versions of the Fay-Herriot model have been proposed. The use of a logarithmic transformation has been studied for example by Slud and Maiti (2006). When the target parameter is a proportion, as is the case for example with the HCR, the use of transformations can be crucial especially when direct estimates are close to the boundary of the support of θ_i. For proportions, the arcsin transformation provides a transformation that can stabilise the variance (Casas-Cordero Valencia, Encina, and Lahiri, 2016). Under this transformation, direct estimates are transformed

$\hat{\theta}_i^{Dir,arcsin} = \sin^{-1}\left(\sqrt{\hat{\theta}_i^{Dir}}\right)$, and the sampling variance, approximated using a Taylor series expansion, is given by $1/(4\tilde{n}_i)$, where $\tilde{n}_i$ is the effective sample size. The Fay-Herriot model is estimated on the transformed scale, and a back-transformation based on the conditional expectation under the model can be employed to account for the non-linearity of the inverse transformation (Sugasawa and Kubokawa, 2017). Other transformations for proportions, including the logit, probit, and complementary log-log, have also been considered in the literature. Jacob et al. (2025) provided a detailed comparison of these alternatives in an application to producing quarterly estimates of extreme poverty rates in Brazil. An alternative framework for area-level modelling of proportions is offered by the extended Beta model proposed by De Nicolò, Fabrizi, and Gardini (2024). Under this approach, the area-level proportions are modelled directly by using a beta distribution, which naturally accounts for the bounded support of the response. The Beta model can be extended to include zero and one-inflated components, which is important in poverty mapping applications, and spatial random effects. This approach has been recently used in a poverty mapping application with remote sensing data, see De Nicolò, Fabrizi, and Gardini (2024). The `tipsae` R package (De Nicolò and Gardini, 2024) provides a Bayesian implementation of the extended Beta models.

In the case of the application in Mozambique, a Fay-Herriot model that uses the arcsin transformation and geospatial predictors aggregated at the district level achieved high correlations with direct estimates at the district and province levels for both small area means and HCRs. These model-based estimates were also comparable to the model-based estimates produced with the Fay-Herriot model that used aggregate predictors from the 2017 census. What was interesting in this case is that using a Fay-Herriot model with district-level predictors coming from the 2007 census produced estimates that were also well correlated with direct estimates. This indicates that area-level aggregation in this case mitigated the effect of possible structural changes in census covariates that, as we described above, can affect the performance of the EBP estimates produced with the unit-level model.

The use of a Fay-Herriot model in the application to data from the UK produced estimates that were substantially better than those under the unit-context model. Combining geospatial covariates with administrative data sources (benefits claimant counts, median house prices, council tax bands, among others) proved to be an effective strategy. This suggests that in some contexts the Fay-Herriot framework can be more effective than unit-context models especially when geospatial data are supplemented with data from other data sources, e.g., administrative data. Initial results from using a Fay-Herriot model with geospatial predictors to data from Greece showed that the resulting estimates are of acceptable quality. Overall, our research evidence so far suggests that area-level models combined with geospatial predictors perform well across applications.

3.4 Variable selection and data quality

When many geospatial layers are available, as is increasingly common with repositories providing several raster products, variable (feature) selection becomes essential. High correlation between geospatial variables can lead to multicollinearity. For example, nighttime lights, road density, and building counts capture aspects of human settlement and economic activity. In addition, as mentioned before, rasters can be aggregated at different spatial scales and by using different summary functions. Regularization methods such as LASSO (Tibshirani, 1996) or elastic net (Zou and Hastie, 2005) provide principled approaches for selecting a parsimonious set of geospatial predictors while addressing multicollinearity. Although moderate multicollinearity may not severely affect SAE point estimates (Merfeld et al., 2023), it can cause numerical instability in model fitting.

Data quality issues also deserve the attention of practitioners. Cloud coverage can introduce noise or missing values in optical satellite imagery. The South Atlantic Anomaly affects the operation of certain satellite instruments, leading to gaps in data coverage over parts of southern Africa and South America. In some cases, specific geospatial products may have incomplete spatial coverage. For example, in the application in Mozambique, we found that Google Open Buildings V3 has missing building footprints in certain regions of the country. This problem was addressed by combining multiple data sources, for example, by combining Google and Microsoft building footprint counts where both are available. Awareness of these limitations is important for practitioners, as the quality of geospatial covariates directly affects the quality of the resulting SAE estimates.

4 Concluding remarks and guidance for practitioners

How to best use geospatial data in model-based small area estimation is application specific. The research evidence from the applications in Mozambique, the UK, and Greece offers valuable insights. In this concluding section, we summarize the main findings.

Unit-level models: Using a well-specified unit-level model and micro-data from a recent census as the source of auxiliary information is the preferred approach for model-based small area estimation. Geospatial data can complement census predictors to assist with reducing the unexplained between-area variability. The main limitation in this case is that the method depends on the availability of population micro-data from a recent census. As illustrated in the case of Mozambique, the use of outdated census data can be risky. Supplementing outdated census data with geospatial data does not guarantee a solution to this problem because census predictors can remain important to achieve a good model fit.

Unit-context models: This model specification requires survey data, the geolocations of the sample units at the chosen spatial scale, population projections at the same spatial scale (if counts from a recent census at the same spatial scale are not available), and corresponding geospatial predictors, making it applicable in settings where recent census microdata are unavailable. How the geospatial predictors are aggregated depends on what information about the geolocations of the sampling units is available, and the type of geospatial variable. As is the case with every model, model-building in the case of the unit-context model requires carefully considering how to include geospatial predictors at different scales. In Mozambique, the use of a unit-context model produced small area estimates of acceptable quality. A finding of practical relevance in this case is that using coarser aggregation levels for geospatial predictors was preferred to using finer-resolution grids. However, the use of the unit-context model in the UK application failed to produce estimates of acceptable quality. This shows that using the unit-context model must be done with great caution.

Area-level models: Area-level models offer a natural framework for integrating survey and geospatial data. Area-level models are deceptively simple. Model specification requires careful consideration of the survey design for point and variance estimation, possible use of variance smoothing methods, and advanced models. Using area-level models with geospatial predictors performed consistently well across the applications we have considered in this paper.

Intercensal updating with geospatial data: Appealing features of geospatial data include global coverage, ease of access, and continuous updating. The last feature is a key reason that has been used to promote the use of geospatial data in SAE applications. Economic development, urbanisation, migration, and policy changes can change the distribution of household characteristics during the intercensal period. A common practical challenge is that, depending on the context, census data

can quickly become outdated. Using outdated census data as predictors relies on the assumption that the relationship between the outcome and the predictors remains stable over time. The data setup in Mozambique allowed us to study the impact of using outdated census data in small area estimation. Comparisons of data from the 2007 and 2017 Mozambique censuses revealed changes in key housing characteristics, including the type of energy source used by households and variables describing the housing quality. These changes reflect a decade of economic development. Using outdated 2007 census data in a unit-level model affected the quality of model-based small area estimates. Instead, the use of a unit-context or an area-level model with geospatial predictors resulted in estimates of acceptable quality. For practitioners working in intercensal periods and without access to population micro-data, these results indicate that geospatial predictors as the source of auxiliary information can be effective. However, a cautious approach to how the model is specified is needed. If analysts are keen to continue using census data, SPREE-type methods (Isidro, Haslett, and Jones, 2016; Luna Hernandez, 2016; Zhang and Chambers, 2004) offer an alternative approach to intercensal updating. Current research focuses on different estimation strategies in the off-census years and shows that SPREE-type methods can be also effective for intercensal updating. Whichever model specification is selected, carefully validating model-based estimates against reliable model-free estimates remains of critical importance.

Software: Processing geospatial data is perhaps the most challenging part for statisticians without access to GIS expertise. Outside specialised GIS software, preprocessing steps to prepare geospatial data for estimation are performed either by using R or Python, both of which now offer mature toolchains for raster and vector operations, as well as access to remote-sensing data. In R, `exactextractr` (Baston, 2023) provides fast zonal statistics for arbitrary polygons, `blackmarbler` (Marty and Stefanini Vicente, 2024) retrieves and processes NASA Black Marble nighttime lights products, and `rgee` (Aybar et al., 2020) exposes the Google Earth Engine catalogue (Gorelick et al., 2017) for Landsat, Sentinel and VIIRS imagery. The recently released `GeoLink` package (Llyod et al., 2025), jointly developed by the World Bank Group and the University of Southampton, provides a higher-level interface that automates the linking of standard geospatial layers, e.g., nighttime lights, building footprints, climate, land cover, among others, to georeferenced survey microdata or shapefiles, considerably reducing the boilerplate involved in assembling SAE-ready covariate sets. Equivalent functionality is available in Python through `rasterio`, `rasterstats` and `geopandas` for raster and vector handling, and through the `earthengine-api` client for Earth Engine access. For small area estimation, the most established implementations of the models discussed here are in R, although Python alternatives are emerging. Among other packages, the `emdi` package (Kreutzmann et al., 2019) provides a unified interface for the area-level and unit-level models, including MSE estimation, as is the `sae` package (Molina and Marhuenda, 2015). The `povmap` package (Edochie, Newhouse, Würz, et al., 2024) extends `emdi` with features specifically tailored to poverty mapping, including a survey-weighted version of EBP, ex-post benchmarking, and a wrapper for use via Stata. For Beta-type models the `tipsae` package (De Nicolò and Gardini, 2024) is the natural choice.

References

Aybar, C., Q. Wu, L. Bautista, R. Yali, and A. Barja (2020). rgee: An R Package for Interacting with Google Earth Engine. In: *Journal of Open Source Software* 5.51, p. 2272. DOI: 10.21105/joss.02272.

Baston, D. (2023). *exactextractr: Fast Extraction of Raster Values Within Polygons*. R package version 0.9.1. URL: https://CRAN.R-project.org/package=exactextractr.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. In: *Journal of the American Statistical Association* 83.401, pp. 28–36. DOI: 10.1080/01621459.1988.10478561.

Casas-Cordero Valencia, C., J. Encina, and P. Lahiri (2016). Poverty Mapping for the Chilean Comunas. In: *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons, Ltd. Chap. 20, pp. 379–404. ISBN: 9781118814963. DOI: https://doi.org/10.1002/9781118814963.ch20. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118814963.ch20. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118814963.ch20.

Chi, G., H. Fang, S. Chatterjee, and E. Blumenstock J. (2022). Microestimates of wealth for all low and middle income countries. In: *PNAS* 119.3, pp. 381–399. DOI: 10.1073/pnas.2113658119.

Corral, P., H. Henderson, and S. Segovia (2025). Poverty Mapping in the Age of Machine Learning. In: *Journal of Development Economics* 172, p. 103377. DOI: 10.1016/j.jdeveco.2024.103377.

De Nicolò, S., E. Fabrizi, and A. Gardini (2024). Extended Beta Models for Poverty Mapping. An Application Integrating Survey and Remote Sensing Data in Bangladesh. In: *The Annals of Applied Statistics* 18.4. DOI: 10.1214/24-AOAS1934.

De Nicolò, S. and A. Gardini (2024). The R Package tipsae: Tools for Mapping Proportions and Indicators on the Unit Interval. In: *Journal of Statistical Software* 108.1, pp. 1–36. DOI: 10.18637/jss.v108.i01.

Edochie, I., D. Newhouse, N. Tzavidis, T. Schmid, E. Foster, A. L. Hernandez, A. Ouedraogo, A. Sanoh, and A. Savadogo (2025). Small Area Estimation of Poverty in Four West African Countries by Integrating Survey and Geospatial Data. In: *Journal of Official Statistics* 41.1, pp. 96–124. DOI: 10.1177/0282423X241284890.

Edochie, I., D. Newhouse, N. Würz, and T. Schmid (2024). *povmap: Extension to the emdi Package*. R package version 1.0.1. DOI: 10.32614/CRAN.package.povmap. URL: https://github.com/SSA-Statistical-Team-Projects/povmap.

Fay, R. E. and R. A. Herriot (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. In: *Journal of the American Statistical Association* 74.366, pp. 269–277. DOI: 10.2307/2286322.

Gelfand, A. E., L. Zhu, and B. P. Carlin (2001). On the Change of Support Problem for Spatio-Temporal Data. In: *Biostatistics* 2.1, pp. 31–45. DOI: 10.1093/biostatistics/2.1.31.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017). Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. In: *Remote Sensing of Environment* 202, pp. 18–27. DOI: 10.1016/j.rse.2017.06.031.

Isidro, M., S. Haslett, and G. Jones (2016). Extended Structure Preserving Estimation (ESPREE) for Updating Small Area Estimates of Poverty. In: *Annals of Applied Statistics* 10.1, pp. 451–76. DOI: 10.1214/15-AOAS900.

Jacob, G. A. P., N. Tzavidis, A. Luna Hernandez, and P. L. D. N. Silva (2025). Quarterly Small Area Estimates of Extreme Poverty in Brazil Using Transformed Fay-Herriot Models. In: *Journal of Survey Statistics and Methodology* 13.5, pp. 552–586. DOI: 10.1093/jssam/smaf025.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019). The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators. In: *Journal of Statistical Software* 91.7, pp. 1–33. DOI: 10.18637/jss.v091.i07.

Llyod, C. T., I. Edochie, D. Jaganjac, J. D. Merfeld, D. Newhouse, L. Perfetti-Villa, D. A. Gomez, and N. Tzavidis (2025). *GeoLink R Package*. DOI: 10.5281/zenodo.15231144. URL: https://doi.org/10.5281/zenodo.15231144.

Luna Hernandez, A. (2016). "Multivariate Structure Preserving Estimation for Population Compositions". PhD thesis. University of Southampton. 127 pp.

Marty, R. and G. Stefanini Vicente (2024). *blackmarbler: Black Marble Data and Statistics*. R package. URL: https://CRAN.R-project.org/package=blackmarbler.

Masaki, T., D. Newhouse, A. R. Silwal, A. Bedada, and R. Engstrom (2022). Small Area Estimation of Non-Monetary Poverty with Geospatial Data. In: *Statistical Journal of the IAOS* 38.3, pp. 1035–1051. DOI: 10.3233/SJI-210902.

Merfeld, J. D., H. Chen, P. Lahiri, and D. Newhouse (2023). Small Area Estimation with Geospatial Data: A Primer. Working Paper.

Molina, I. and Y. Marhuenda (2015). sae: An R Package for Small Area Estimation. In: *The R Journal* 7.1, pp. 81–98. DOI: 10.32614/RJ-2015-007.

Molina, I. and J. N. K. Rao (2010). Small Area Estimation of Poverty Indicators. In: *Canadian Journal of Statistics* 38.3, pp. 369–385. DOI: 10.1002/cjs.10051.

Newhouse, D., A. Ramakrishnan, T. Swartz, J. Merfeld, and P. Lahiri (2025). Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning. In: *Oxford Bulletin of Economics and Statistics*. DOI: 10.1111/obes.12678.

Nguyen, V. C. (Dec. 2012). A Method to Update Poverty Maps. In: *Journal of Development Studies* 48.12, pp. 1844–1863. DOI: 10.1080/00220388.2012.682983.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2019). Data-Driven Transformations in Small Area Estimation. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 183.1, pp. 121–148. DOI: 10.1111/rssa.12488.

Slud, E. V. and T. Maiti (2006). Mean-Squared Error Estimation in Transformed Fay–Herriot Models. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.2, pp. 239–257. DOI: 10.1111/j.1467-9868.2006.00542.x.

Stevens, F., A. Gaughan, C. Linard, and A. Tatem (2017). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. In: *PLoS ONE* 10(2): e0107042. DOI: 10.1371/journal.pone.0107042.

Sugasawa, S. and T. Kubokawa (2017). Transforming Response Values in Small Area Prediction. In: *Computational Statistics & Data Analysis* 114, pp. 47–60. DOI: 10.1016/j.csda.2017.03.017.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Trevisani, M. and A. Gelfand (2013). "Spatial Misalignment Models for Small Area Estimation: A Simulation Study". In: *Advances in Theoretical and Applied Statistics*. Studies in Theoretical and Applied Statistics. Berlin, Heidelberg: Springer, pp. 269–279. DOI: 10.1007/978-3-642-35588-2_25.

Van der Weide, R., B. Blankespoor, C. Elbers, and P. Lanjouw (2022). How Accurate Is a Poverty Map Based on Remote Sensing Data? An Application to Malawi. World Bank Policy Research Working Paper. URL: https://openknowledge.worldbank.org/server/api/core/bitstreams/b6d07e8a-7a03-58fe-b8f1-2ac86fa12011/content.

Zhang, L.-C. and R. Chambers (2004). Small area estimates for cross□classifications. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.2, pp. 479–496. DOI: 10.1111/j.1369-7412.2004.05266.x.

Zou, H. and T. Hastie (2005). Regularization and Variable Selection Via the Elastic Net. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, pp. 301–320.