Extended Beta models for poverty mapping. An application integrating survey and remote sensing data in Bangladesh

Submitted to the Annals of Applied Statistics

EXTENDED BETA MODELS FOR POVERTY MAPPING. AN APPLICATION INTEGRATING SURVEY AND REMOTE SENSING DATA IN BANGLADESH

BY SILVIA DE NICOLÒ[1,a], ENRICO FABRIZI[2,c] AND ALDO GARDINI[1,b]

[1]*Department of Statistical Sciences, Università di Bologna,* [a]*silvia.denicolo@unibo.it;* [b]*aldo.gardini@unibo.it*

[2]*DISES, Università Cattolica del S. Cuore ,* [c]*enrico.fabrizi@unicatt.it*

The paper targets the estimation of a poverty rate at the upazila level in Bangladesh through the use of Demographic and Health Survey (DHS) data. Upazilas are administrative regions equivalent to counties or boroughs whose sample sizes are not large enough to provide reliable estimates or are even absent. We tackle this issue by proposing a small area estimation model complementing survey data with remote sensing information at the area level. We specify an Extended Beta mixed regression model within the Bayesian framework, allowing it to accommodate the peculiarities of sample data and to predict out-of-sample rates. Specifically, it enables to include estimates equal to either 0 or 1 and to model the strong intra-cluster correlation. We aim at proposing a method that can be implemented by statistical offices as a routine. In this spirit, we consider a regularizing prior for coefficients rather than a model selection approach, to deal with a large number of auxiliary variables. We compare our methods with existing alternatives using a design-based simulation exercise and illustrate its potential with the motivating application.

1. Introduction. There is a growing interest in the study of geographical distribution of extreme poverty, with a particular focus on developing countries, due to the relevance of place-based policies implementation and monitoring (Duranton and Venables, 2021). In most countries, the parameters usually adopted to describe poverty and social exclusion are estimated using sample surveys, providing reliable estimates for the country as a whole, for large regions, or for other large subsets of the population. Nonetheless, the availability of estimates for small geographical regions or other small subsets of the population, usually labeled as *small areas* or *domains*, is particularly useful. When the domain-specific sample sizes are too small, the precision of survey estimates is not adequate. Small area estimation (SAE) methods aim at improving the precision of area-specific survey estimates (known as *direct* estimates) by integrating survey samples with different data sources that can provide indirect useful information. In this article, we aim at mapping poverty in Bangladesh at a great level of disaggregation using data from the Bangladesh Demographic and Health Surveys (DHS). Specifically, we consider as target areas the upazilas, i.e. administrative sub-districts comparable with counties or boroughs.

As poverty measure, we consider the proportion of people in the first quintile of the national distribution of the Wealth Index (WI), as defined by the DHS program (Corsi et al., 2012). The WI is a composite measure that summarizes the living conditions of an household and can be read as a measure of socioeconomic status (Poirier, Grépin and Grignon, 2020). Such indicator is more closely related to permanent than to current income, being less reactive to changes in income or consumption than other poverty measures, as noted by Steele et al. (2017) for the Bangladesh case. We remark that surveys implemented by the

Keywords and phrases: Demographic Health Survey, Hierarchical Bayes, Shrinkage priors, Small area estimation.

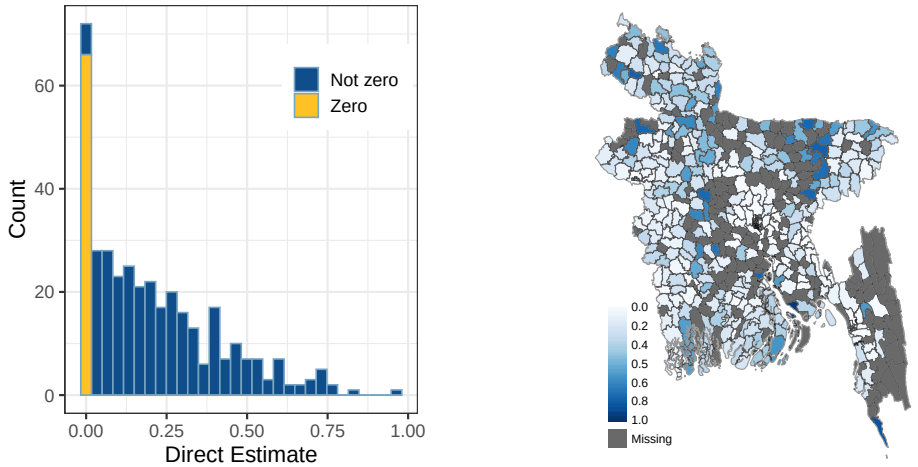

FIG 1. *Direct estimates of poverty incidence in Bangladeshi Upazilas from DHS survey: histogram (left) and map (right).*

DHS program constitute a valuable data source, being collected with similar methodologies in many developing countries.

Due to the lack of reliable and standardized data sources released by national institutions, the DHS program promotes the incorporation of geo-referenced data (Burgert, Zachary and Colston, 2013). In this spirit, we integrate auxiliary information taken from remote sensing (RS), such as population structure and density, along with geographical, land use, social and economic features. Those are largely recognized as poverty predictors and used for poverty estimation at fine spatial levels (Engstrom, Hersh and Newhouse, 2017; Masaki et al., 2020). In Bangladesh, some poverty-related measures have been predicted at small scales by integrating survey and RS data via novel statistical techniques (Steele et al., 2017; Zhao et al., 2019). The approaches followed by these papers, however, completely overlook the uncertainty of survey estimates used as input. Such uncertainty might be non-negligible in small domains and ignoring it may lead to unreliable predictions and misleading associated variances. For these reasons, we decided to exploit SAE techniques, leveraging sampling variances of survey estimators to obtain reliable estimates endowed with adequate precision measures.

To produce a reliable map at the upazila level, we need to face several specific challenges posed by our data set: *(a)* samples available for upazilas are often very small and, for more than 30% of the areas, no observations are recorded (see spatial distribution of direct estimates in Figure 1); *(b)* the sampling design is clustered and the intra-cluster correlation is very high (see Section 3, for details), so direct estimators are much less efficient than they would be under a simple random sampling design with equal sample sizes; *(c)* Bangladesh is characterized by large disparities between urban and rural areas and among regions: as a consequence a large fraction, 18%, of direct estimates are 0, as we can note from the histogram in Figure 1; *(d)* the auxiliary information we have is areal and although in most cases is available at a very detailed scale, it cannot be linked to individuals as exact geographical coordinates of sampled households are not released; *(e)* the potentially large number of covariates that can be obtained for each area poses a problem of variable selection. These challenges drive our modeling choices and lead us to develop a methodology that is fairly different from those already proposed in the literature.

We consider small area methods relying on area-level models (Rao and Molina, 2015, Chapter 4). The alternative unit-level models (see, e.g., Molina, Nandram and Rao, 2014)

require auxiliary variables to be known for each unit or household in the sample and also for all the non-sampled units, in case of prediction. An appealing feature of area-level models is their ability to link area-level covariates with direct estimates of the target measure, endowed with an uncertainty measure. They do not require access to individual-level sample data and detailed design information, often not available because of confidentiality concerns. Among area-level models for proportions, the main proposals are linear mixed models assuming normality (i.e., the Fay-Herriot model, FH, Fay and Herriot, 1979) either on direct estimates (Xie, Raghunathan and Lepkowski, 2007; Marhuenda, Molina and Morales, 2013) or on their transformations, e.g. the square root arc-sine one (Casas-Cordero Valencia, Encina and Lahiri, 2016; Schmid et al., 2017), and Beta mixed models (Liu, Lahiri and Kalton, 2007; Fabrizi et al., 2011; Janicki, 2020). Other proposals directly model survey counts through discrete distributions such as the Binomial (Chen, Wakefield and Lumely, 2014; Franco and Bell, 2015) or the Poisson (Bradley, Wikle and Holan, 2016; Boubeta, Lombardía and Morales, 2017).

We remark that the assumption of normality for direct estimators works only when the underlying parameter is far from the boundaries. Likewise, the square root Arc-sine transformation model (hereafter Arc-sine model, for brevity) implies an approximation of standard errors that is inadequate for direct estimates equal to 0 or very close to it (see the discussion in Section 6.1). When estimates are far from 0 or 1, the problem can be overlooked, but this is not our case. Lastly, Poisson and Binomial models do not present a dispersion parameter, complicating the incorporation of sampling variances and intra-cluster correlation.

In view of the considerations stated in the previous paragraph, this paper focuses on Beta mixed models. Indeed, despite Beta models do not allow for direct estimates equal to 0 or 1, ad-hoc solutions for this issue have been proposed in SAE by Wieczorek and Hawala (2011), Fabrizi, Ferrante and Trivisano (2016) and Fabrizi, Ferrante and Trivisano (2020). However, these contributions do not consider the strong intra-cluster correlation characterizing our data and may lead to poor predictions for these areas. In this paper, we opt for an extended Beta mixed model that we generalize to make it suitable for the aforementioned features. The inferential setting we adopt is the Bayesian one, as it offers several benefits in the small area estimation context (Rao and Molina, 2015, Chapter 10). Among the others, we point out those of easily managing non-Gaussian distributional assumptions and fully capturing the uncertainty around target parameters.

Our proposal contributes to the literature in various directions. Firstly, we substantially extend the model of Fabrizi, Ferrante and Trivisano (2016) by proposing a different treatment of the direct estimates equal to the boundaries. We assume that either 0 or 1 values are due to a censoring process as a result of reduced area-specific sample sizes and strong intra-cluster correlation, while true population values may be very close but not exactly equal to the boundary values. Differently from Fabrizi, Ferrante and Trivisano (2016), we relax the independence assumption when modeling the probability of observing 0 and 1 values. We explicitly include an additional parameter that manages such dependency in an intuitive and explicable way. Moreover, we propose a new methodology for out-of-sample predictions that properly represent the associated uncertainty. Specifically, it guarantees that this uncertainty will be larger than the one associated with areas for which samples are available. This may not be the case for many small area procedures in use.

Another goal of the proposed method is to provide a small area estimation tool that can be widely applied by final users and practitioners. We achieve this by implementing a set of flexible prior choices that do not need fine-tuning interventions. In this spirit, the model selection step is automatically performed by using regularized horseshoe priors (Piironen and Vehtari, 2017) for regression coefficients, sidestepping manual variables selection and dimension reduction techniques. To the best of our knowledge, this constitutes a novelty in

the small area framework. Moreover, the use of the horseshoe prior for regression coefficients is complemented by a type of spike-and-slab prior for the random effects (Fabrizi, Ferrante and Trivisano, 2018; Tang et al., 2018).

We assess the frequentist properties of the proposed predictor using a design-based simulation in comparison with existing models in the literature, namely the Arc-sine model, the model by Fabrizi, Ferrante and Trivisano (2016) among those relying on the Beta likelihood, and the Binomial model. We do not consider other modeling proposals such as those based on a Poisson working likelihoods for several reasons. For example, they do not naturally restrict the domain of the target (rate) parameter within the unit interval and, from a technical point of view, their implementation to fit our data is not straightforward and would require an out-of-the-scope investigation for this paper. A detailed discussion can be found in Section 7.

We find that the predictors we propose are very effective in improving the precision of direct estimates, having good coverage properties in terms of posterior probability intervals for both in-sample and out-of-sample areas.

The paper is organized as follows. In Section 2, the DHS survey and auxiliary variables are presented. The direct estimation of proportions is set out in Section 3, together with a particular focus on the methodology of uncertainty estimation that has been adopted. The small area models are introduced in Section 4, deepening the proposed Extended Beta model. Section 5 deals with a design-based simulation study, whereas an application on Bangladesh DHS data is illustrated in Section 6. Section 7 offers some concluding remarks and directions for future research.

2. The data. We aim at estimating the proportion of people living in households below the 20th percentile of the WI national distribution at the Administrative Level 3 (upazilas) in Bangladesh. To obtain these estimates, we combine DHS survey data, described in Section 2.1, with remote sensing and geographical data (Section 2.2) obtained from a variety of open sources and processed at the upazila level.

2.1. *The Bangladesh DHS.* The DHS survey targets the entire Bangladeshi population residing in non-institutional dwelling units. Bangladesh is divided into seven administrative divisions; each division into zilas and each zila into upazilas. The national territory is also classified distinguishing among rural areas, city corporations and other urban areas. The survey is based on a two-stage stratified sample of households and relates to 2014. In the first stage, 600 Enumeration areas (EAs) are selected with probability proportional to the EA size within 20 strata, obtained as the combination of administrative divisions and territorial classification (originally, 21 strata were planned, but the city corporation and other urban areas of the Rangpur division were merged). Each EA is defined to contain on average 120 households, and 30 households are drawn from every sampled EA with equal probability. Of the 600 EAs in the sample, 207 are in urban areas or city corporations and the remaining 393 are in rural areas. With this design, the survey selects 18,000 residential households. Survey weights, accounting also for non-responses, are published with survey data (NIPORT and Mitra and Associates and ICF International, 2016, for more details). We have 365 upazilas that include at least one sampled cluster out of a total of 544.

The WI computed using DHS data is constructed from a set of questions on household durable assets and housing characteristics such as floor type and ceiling material, toilet or latrine access, phone ownership and others. Given the set of basic indicators, the construction of the index proceeds by extracting a common factor explaining the largest percentage of the total variance using principal component analysis and then adjusting for differences in urban and rural strata. Households with a WI included in the first quintile are labeled as *poor*, defining a dichotomous response variable denoted with y. In line with literature on poverty

measurement, we target our analysis at the individual level: as a consequence, all individuals belonging to the same household are assumed to share the same WI score. The individual data is characterized by an overall sample size of 81,624, while the upazila-level sample sizes span from 16 to 1884 (median: 160).

2.2. *Remote sensing covariates.* According to World Bank (2008), Khudri et al. (2013) and Islam, Sayeed and Hossain (2017), the main determinants of poverty relate to socio-demographic and educational aspects, economic development and the so-called "location effect". The latter is associated with connectivity to markets and infrastructures (for rural communities), area-specific risk of natural disasters and lean seasons related to area-specific crops. With the exclusion of the education level, not considered due to the non-availability of data, we incorporate all those aspects through selected covariates, described in the following. In particular, location-specific issues have been captured with the aid of land-use and bio-climatic variables.

We have chosen a set of auxiliary variables, aggregated at the area level from raster files available in different open sources. A total of $p = 46$ covariates are included in the application. All the covariate values at the area level are retrieved by cutting the raster with the Bangladesh upazila shapefile first, and then simply aggregating the pixels inside each area. Usually, the arithmetic mean is considered, but different summaries, that are meaningful for specific indicators, are described later. We remark that this procedure allows producing area level covariates starting from open resources. To improve the strength of the linear correlation between each covariate and the transformed proportion (i.e. logit or arc-sine), some data transformations are considered (identity, logarithm, squared root and inverse functions) choosing the one that maximizes Pearson's correlation. Lastly, the obtained covariates are standardized. The raster related to population density has a resolution of approximately one pixel per km^2, while the others have a resolution of approximately one pixel per hectare.

2.2.1. *Demographic variables.* The demographic structure of the areas is described by the population density and its composition by age and sex, retrieved from the rasters available on the WorldPop website (https://www.worldpop.org/, Tatem, 2017). Regarding the density, the estimate of the count of People-per-km^2 is available and it has been summarized in each area by the average and the standard deviation. On the other hand, the population structure by age and sex is available as rasters reporting the counts of People-per-hectare, for each of the following age classes: $[0; 1)$, $[1; 5)$, $[5; 10)$, $[10; 15)$, $\ldots$, $[75; 80)$, $[80; +\infty)$, and stratified by gender (see Pezzulo et al., 2017, for the methodology). Let us define $P_{G,A}$ as the population count pertaining to gender G and age class A. By summing each count within the target administrative areas, we produce the following demographic ratios: human sex ratio $P_{M,\bullet}/P_{F,\bullet}$, human sex ratio in productive age $P_{M,14-64}/P_{F,15-64}$, total dependency ratio, i.e., $(P_{\bullet,0-14} + P_{\bullet,65+})/P_{\bullet,15-64}$, child dependency ratio, i.e., $P_{\bullet,0-14}/P_{\bullet,15-64}$, aged dependency ratio, i.e., $P_{\bullet,65+}/P_{\bullet,15-64}$ and woman-child ratio $P_{F,15-49}/P_{\bullet,0-4}$.

2.2.2. *Development variables.* As an indicator of the area urbanization, the nighttime light radiance (from WorldPop) is adopted, measured by Visible Infrared Imaging Radiometer Suit (VIIRS, nanoWatts/cm^2/sr) and is acknowledged to be a proxy of economic development (Zhou et al., 2015; Masaki et al., 2020). Further information on the development of an area is retrieved from the distances to facilities and main infrastructures. More in detail, we considered the distance in km to important road intersections, roads, waterways (from WorldPop) and the time required to access the city and the nearest healthcare site, coming from the Malaria Atlas Project (https://malariaatlas.org/explorer/, Hay and Snow, 2006). Note that, since these quantities are strictly related to people living in the area, the average

was computed by weighting each pixel with the corresponding population density. To do this, the rasters with a resolution of one hectare need to be up-scaled and aligned to the raster of the population density.

2.2.3. *Land-use variables.* Another important aspect to take into account is the kind of use that a territory has. To this aim, we consider again rasters from WorldPop, including the average distance of each pixel from areas with a determined classification of use (cultivated, woody-tree, shrub, herbaceous, sparse vegetation, aquatic vegetation, artificial surface, bare area, nature reserves, open-water coastline). To complete the physical characterization of the territory, the elevation above sea level and the topographic slope are averaged within each area.

2.2.4. *Bio-climatic variables.* Such covariates are useful to account for the weather conditions that affect the areas. They constitute a set of 19 variables, available in the WorldClim repository (https://www.worldclim.org/data/bioclim.html, O'Donnell and Ignizio, 2012) that is built in order to summarise the overall and seasonal behaviors of temperature and rainfall (e.g., annual mean, standard deviation and temperature diurnal range). The available rasters contain the averaged values over the period 1970-2000, providing a static characterization of climatic features. However, given that the agricultural sector employs a large fraction of the workforce in Bangladesh and constitutes a driving force for its economic growth (Rahman et al., 2017), such features may be helpful in characterizing the productivity of the area.

3. Poverty estimator. In this section, we introduce the direct estimator $\hat{\bar{Y}}_d$ of the headcount poverty rate θ_d for the upazila d, based on a complex survey sample of n_d individuals clustered in m_d households. The individual sample size is obtained as $n_d = \sum_{h=1}^{m_d} k_{dh}$ where k_{dh} is the number of components in household h in area d. The estimator also considers the sample weights w_{dh} and the value of the target variable y_{dh}, i.e. an indicator variable denoting the poverty status. We employ an Hájek-type estimator (Hájek, 1971) defined as

$$\hat{\bar{Y}}_d = \frac{\sum_{h=1}^{m_d} k_{dh} w_{dh} y_{dh}}{\sum_{h=1}^{m_d} k_{dh} w_{dh}}, \qquad d \in 1, \dots D_s \tag{1}$$

with D_s being the number of in-sample upazilas. The proportion estimator $\hat{\bar{Y}}_d$ is asymptotically unbiased.

3.1. *Uncertainty associated to direct estimators.* The small-area models that we are going to discuss in Section 4 require a dispersion parameter that can be expressed as a function of the effective individual sample size $\tilde{n}_d$. Such quantity corresponds to the virtual size of a simple random sample producing a direct estimate with a standard error equal to the one obtained under the actual design. It can be characterized as $\tilde{n}_d = n_d/\text{DEff}_d$ where DEff_d denotes the design effect, i.e. the ratio between the design-based variance of a generic estimator and the simple random sampling variance. It measures the possible amount of variance inflation induced by clustering caused by the complex selection process and has to be estimated.

In principle, the sampling variance of (1) under complex two-stage sampling designs is estimated through the Ultimate Cluster technique (Kalton, 1979), where variability among clusters plays a central role. In practice, for many areas, such estimates are unstable or even impossible to be obtained as a low number of clusters (often only one) is available. To circumvent this problem, we obtain reliable estimates of design effects at a higher level of aggregation, subsequently assigning them at the upazila level.

Specifically, our proposal is to consider the 21 strata of the sampling design to estimate the design effect for each stratum $s = 1, \dots, 21$. At the stratum level, the features to be accounted

Stratum Type	Average $\sqrt{\text{DEff}_s}$	Average ρ_s
Rural	4.75	0.18
Other Urban	6.07	0.28
City Corp.	2.93	0.02

TABLE 1

Arithmetic mean of DEff$_s$ and harmonic mean of ρ_s within strata types.

for in the computation of the design effects are the unequal sampling weights and clustering (Chen and Rust, 2017). For this reason, we decide to adopt the formula by Kish (1987) within each stratum, blending weights and clustering components. The formula has been adapted by Gabler, Häder and Lahiri (1999) and adjusted by Lynn, Häder and Gabler (2006). It is defined as

$$\text{DEff}_s = \left[1 + \text{cv}^2(\mathbf{w}_s)\right]\left[1 + (n_s^* - 1)\rho_s\right], \tag{2}$$

where $\text{cv}(\mathbf{w}_s)$ is the coefficient of variation of the vector $\mathbf{w}_s$ of weights associated with individuals in stratum s, inheriting the weight from the household they belong to; ρ_s is the intra-cluster correlation coefficient and

$$n_s^* = \frac{\sum_{i=1}^{c_s}\left(\sum_{j=1}^{n_i} w_{ij}\right)^2}{\sum_{i=1}^{c_s}\sum_{j=1}^{n_i} w_{ij}^2},$$

with c_s being the number of clusters in s, n_i the units within cluster i, and w_{ij} the individual weight.

The intra-cluster correlation coefficient is estimated through an ANOVA-based estimator among those proposed by Ridout, Demetrio and Firth (1999), suitable for the analysis of binary data. Table 1 summarizes the main results of the estimation of DEff$_s$ in different types of strata according to the habitation type (see Section 2.1). *Rural* and *Other Urban* strata show particularly high ICCs and, consequently, high estimates of DEff$_s$. On the other hand, *City Corp.* strata have lower design effects in view of their lower ICCs. In three *City Corp.* strata, ρ_s cannot be computed due to the absence of poor households in the observed sample: in these cases, we impute the harmonic mean of ICCs pertaining to *City Corp.* strata (see Table 1). Once the DEffs are available, standard errors are computed starting from the definition of variance of proportions under simple random sampling:

$$\hat{\mathbb{SE}}_{cs}\left[\hat{\bar{Y}}_d\right] = \sqrt{\frac{\hat{\bar{Y}}_d(1-\hat{\bar{Y}}_d)}{n_d}\text{DEff}_s}. \tag{3}$$

To validate the procedure, we remark that the linear correlation between standard errors as in (3) and the Ultimate Cluster estimates at the strata level is 0.93. At this level, the Ultimate Cluster technique is reliable due to the large number of clusters in each stratum. This comparison shows that both strategies provide similar results, leading us to consider DEff estimates as reliable.

4. The models. The model strategy we propose, which constitutes an extension of the one proposed by Fabrizi, Ferrante and Trivisano (2016), is fully described in Section 4.1. An alternative approach relying on the Arc-sine model (Schmid et al., 2017) is presented in Section 4.2, whereas Section 4.3 the approach introduced by Chen, Wakefield and Lumely (2014) that assumes a Binomial working likelihood.

4.1. *The Extended Beta Model.* Let us consider the mean-precision parametrization of the Beta random variable (Ferrari and Cribari-Neto, 2004): if $Y \sim \text{Beta}(\mu\phi, (1-\mu)\phi)$, then

$$f_B(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\,\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad y \in (0,1),$$

where $\mu \in (0,1)$ is the expectation and $\phi \in (0,+\infty)$ is a dispersion parameter as $\mathbb{V}[y] = \mu(1-\mu)(\phi+1)^{-1}$. Note that the expression of $\mathbb{V}[y]$ is consistent with equation (3), leading to an intuitive interpretation of the Beta distribution in modeling proportions. As a consequence, $\phi+1$ can be interpreted as the equivalent sample size. In SAE context, the Beta regression area level model (Janicki, 2020) is usually specified as

$$\begin{aligned} &\hat{\bar{Y}}_d|\mu_d,\phi_d \overset{ind}{\sim} \text{Beta}\left(\mu_d\phi_d, (1-\mu_d)\phi_d\right), \\ &\text{logit}\left(\mu_d\right) = \mathbf{x}_d^T\boldsymbol{\beta} + v_d, \quad d = 1,\dots,D, \end{aligned}$$

with $\mathbf{x}_d$ being a set of p covariates, $\boldsymbol{\beta}$ the vector of regression coefficients, v_d a random effect and ϕ_d the area-specific dispersion parameter.

In order to allow direct estimates to be equal to 0 and 1, the standard Beta model has to be extended. We start by considering the following three-components mixture model, consistently with Wieczorek and Hawala (2011):

$$\begin{aligned} &\hat{\bar{Y}}_d|\mu_d,\pi_{0d},\pi_{1d} \overset{ind}{\sim} \pi_{0d} \times \mathbf{1}\{\hat{\bar{Y}}_d = 0\} + \\ &\qquad + (1-\pi_{0d}-\pi_{1d}) \times \text{Beta}\left(\mu_d\phi_d, (1-\mu_d)\phi_d\right) \mathbf{1}\{\hat{\bar{Y}}_d \in (0,1)\} + \\ &\qquad + \pi_{1d} \times \mathbf{1}\{\hat{\bar{Y}}_d = 1\},\ d = 1,\dots,D \\ &\text{logit}\left(\mu_d\right) = \mathbf{x}_d^T\boldsymbol{\beta} + v_d. \end{aligned} \tag{4}$$

with π_{0d} and π_{1d} denoting the probabilities of observing 0 and 1 values in area d. The way we model such probabilities is the main point of divergence with Wieczorek and Hawala (2011): while they define π_{0d} and π_{1d} as the result of two logistic regressions, requiring a reasonable amount of information, we decide to adopt a more parsimonious approach. In this way, our model can be estimated even when boundary values are sparse.

The basic idea is to assume that possible direct estimates equal to 0 or 1 are the output of a censoring process, i.e. the actual population value θ_d cannot be exactly 0 or 1. This assumption leads to the following definition of the parameters π_{0d} and π_{1d}:

$$\pi_{0d} = \mathbb{P}[\hat{\bar{Y}}_d = 0|\theta_d \in (0,1)], \quad \pi_{1d} = \mathbb{P}[\hat{\bar{Y}}_d = 1|\theta_d \in (0,1)].$$

To express them in a parsimonious way, we decided to define them as a combination of sample characteristics and probabilistic assumptions.

Let us recall that estimator (1) is based on the sequence of observation $y_{d1},\dots,y_{dm_d}$ denoting the household poverty status. This can be seen as a sequence of Bernoulli trials with a probability of success $\mathbb{P}[y_{dh} = 1|\theta_d \in (0,1)] = \mu_d$, $\forall h$, since μ_d may be seen as the poverty rate of non-censored observations. Such an approach for modeling π_{0d} and π_{1d} resembles the one of Fabrizi, Ferrante and Trivisano (2016), but it extends it in different ways. First, we introduce the possibility of observing also direct estimates equal to 1. Secondly, we relax their assumptions of independence across household observations, which is inconsistent with the evidence of a strong clustering effect.

The sequence of observations incorporates a complex dependency structure which results to be challenging to model. For this reason, we opt for a simple and general assumption:

the dependency across observations boils down to a pairwise dependency, which is constant across pairs and areas, not depending on their order, namely

$$\mathbb{P}[y_{di}=1|y_{d1}=1,\dots,y_{d(i-1)}=1,y_{d(i+1)}=1,\dots,y_{dm_d}=1]=\mathbb{P}[y_{di}=1|y_{dh}=1]=\lambda,$$

where $h\neq i$ picks a generic observation. This assumption can be seen as a generalization of Markov dependence in which the ordering does not play a role and allows for exchangeability of the conditional probabilities. In this context, following Klotz (1973), we can formalize π_{1d} as

$$\pi_{1d}=\mathbb{P}[y_{d1}=1,\dots,y_{dm_d}=1|\theta_d\in(0,1)]=\mu_d\lambda^{m_d-1}, \tag{5}$$

i.e. the probability of jointly observing a sequence of m_d ones. Furthermore, in view of

$$\mathbb{P}[y_{di}=0|y_{dh}=0,\theta_d\in(0,1)]=\frac{1+\mu_d(\lambda-2)}{1-\mu_d},$$

we can also define

$$\pi_{0d}=\mathbb{P}[y_{d1}=0,\dots,y_{dm_d}=0|\theta_d\in(0,1)]=\frac{[1+\mu_d(\lambda-2)]^{m_d-1}}{(1-\mu_d)^{m_d-2}}. \tag{6}$$

Note that the additional parameter λ can be interpreted as a proxy of the correlation between household observations and has a bounded support:

$$\lambda_L=\max\left\{0,\max_d\frac{2\mu_d-1}{\mu_d}\right\}\leq\lambda\leq 1.$$

For a specific area d, if $\mu_d<\lambda\leq 1$ holds, a positive correlation across observations is present since observing a success makes more likely the occurrence of another success. On the other hand, $\lambda_L\leq\lambda<\mu_d$ implies a negative correlation, while $\lambda=\mu_d$ implies no correlation. In the latter case, note that $\pi_{0d}=(1-\mu_d)^{m_d}$ and $\pi_{1d}=\mu_d^{m_d}$ as in Fabrizi, Ferrante and Trivisano (2016). Generally speaking, λ has an interpretation also when the pairwise dependency assumptions are relaxed. In this case, π_{1d} can be written as:

$$\begin{aligned}\mu_d\lambda^{m_d-1}&=\mathbb{P}[y_{d1}=1,\dots,y_{dm_d}=1|\theta_d\in(0,1)]\\&=\mathbb{P}[y_{d1}=1|\theta_d\in(0,1)]\prod_{i=2}^{m_d}\mathbb{P}[y_{di}=1|y_{d(i-1)}=1,\dots,y_{d1}=1,\theta_d\in(0,1)],\end{aligned}$$

leading to

$$\lambda=\left(\prod_{i=2}^{m_d}\mathbb{P}[y_{di}=1|y_{d(i-1)}=1,\dots,y_{d1}=1,\theta_d\in(0,1)]\right)^{\frac{1}{m_d-1}}.$$

As a consequence, the additional parameter λ is nothing more than the geometric mean of the m_d-1 conditional probabilities of success, assumed to be constant across areas. Indeed, we explicitly incorporate dependency in a simple and easy-to-interpret way.

Under model (4) and relations (5) and (6), it is possible to express the population proportion θ_d in terms of λ as

$$\theta_d=\left[1-\mu_d\lambda^{m_d-1}-\frac{[1+\mu_d(\lambda-2)]^{m_d-1}}{(1-\mu_d)^{m_d-2}}\right]\mu_d+\mu_d\lambda^{m_d-1}. \tag{7}$$

This implies that θ_d depends on the (inverse logit-transformed) linear predictor itself, which is updated with sample features and λ, a parameter describing the censoring process. Lastly,

the conditional variance is defined as

$$\mathbb{V}\left[\hat{\tilde{Y}}_d|\mu_d,\pi_{0d},\pi_{1d}\right] = (1-\pi_{0d}-\pi_{1d})\,\frac{\mu_d(1-\mu_d)}{\phi_d+1} + \pi_{1d}(1-\pi_{1d}) + \\ + (1-\pi_{0d}-\pi_{1d})\,\mu_d^2\left[\pi_{0d}+\pi_{1d}-2\frac{\pi_{1d}}{\mu_d}\right]. \tag{8}$$

Before we turn to prior specification, we note that the parameter ϕ_d is assumed known, in line with many small area estimation applications, to guarantee identifiability. For this reason, in what follows, it will be replaced by $F_d = \tilde{n}_d - 1$ that is intuitively grounded in the interpretation of the re-parametrized Beta. The effective sample size $\tilde{n}_d$ is estimated as in Subsection 3.1. Note from (8) that F_d appears in $\mathbb{V}\left[\hat{\tilde{Y}}_d|\mu_d,\pi_{0d},\pi_{1d}\right]$ only in the first addend, that is related to the occurrence of $\hat{\tilde{Y}}_d \in (0,1)$.

4.1.1. *Prior specification.* The following prior distributions complete the model. Let us start from λ, for which we opt for a non-informative approach by adopting a Uniform distribution on its support:

$$\lambda|\mu_1,\ldots,\mu_D \sim \text{Unif}\left[\max\left\{0,\max_d \frac{2\mu_d-1}{\mu_d}\right\};1\right].$$

As regards the regression slopes, since we are dealing with a very large number of covariates, a shrinking prior on regression coefficients may be appealing to regularize the problem and avoid a formal step of variable selection or reduction of the predictors space. Specifically, the regularized horseshoe prior proposed by Piironen and Vehtari (2017) is considered, whose basic rationale is that of coercing to 0 the coefficients related to negligible covariates. It is defined by the following mixture:

$$\begin{aligned} &\beta_j|\zeta_j,\tau,\iota \sim \mathcal{N}\left(0,\tau^2\tilde{\zeta}_j^2\right),\ \tilde{\zeta}_j^{\,2} = \frac{\iota^2\zeta_j^2}{\iota^2+\tau^2\zeta_j^2},\ j=1,\ldots,p;\\ &\zeta_j \sim \text{Half-Cauchy}(0,1),\ j=1,\ldots,p;\\ &\iota^2 \sim \text{Inverse-Gamma}\left(\frac{\nu_{slab}}{2},\frac{\nu_{slab}}{2}s_{slab}^2\right);\\ &\tau \sim \text{Half-Cauchy}(0,\tau_0). \end{aligned} \tag{9}$$

In order to complete the prior specification, some hyperparameters need to be set: ν_{slab} and s_{slab} can be interpreted, respectively, as the degrees of freedom and scale of a Student's t prior assumed on coefficients far from zero. We decided to set $s_{slab}=1$, $\nu_{slab}=5$, in order to facilitate the convergence of the MCMC algorithm. Eventually, τ_0 represents an important parameter to set; Piironen and Vehtari (2017) proposed the following expression:

$$\tau_0 = \frac{p_0\tilde{\sigma}}{(p-p_0)\sqrt{D}},$$

where p_0 is an initial guess of the number of non-zero coefficients (i.e. specific of the application) and $\tilde{\sigma}^2$ is the pseudo-variance of a generic observation under the assumed model. To elicit a value for $\tilde{\sigma}^2$ under the Beta model, we exploit a result by Ferrari and Cribari-Neto (2004). They define the logit transformations of the responses: $\mathbf{z} = \{\text{logit}(\hat{\tilde{Y}}_d)\}$, $d \in D_s$ and note that, under the logit link, the unconditional variance of the data can be approximated by:

$$\tilde{\sigma}^2 = \frac{\sum_{d\in D_s}(z_d-\bar{z})^2}{D_s-1}\frac{1}{\bar{\mu}^2(1-\bar{\mu})^2}, \tag{10}$$

where

$$\bar{\mu} = \frac{e^{\bar{z}}}{1+e^{\bar{z}}}.$$

When direct estimates are very imprecise and/or the predictive power of predictors is relevant, most of the random effects can be very small with possibly few exceptions (Datta, Hall and Mandal, 2011). In this line, we propose the variance gamma shrinkage prior introduced by Brown and Griffin (2010) and implemented in a small area application by Fabrizi, Ferrante and Trivisano (2018) as a prior choice for v_d. It is a global-local shrinkage prior also mentioned among those explored by Tang et al. (2018), enabling for shrinking to 0 the random effects related to a subset of the areas by mimicking the behavior of a spike-and-slab prior. More in detail, we specify:

$$\begin{aligned} v_d|\psi_d,\xi &\overset{ind}{\sim} \mathcal{N}\left(0,\psi_d\xi^2\right),\ d=1,\dots,D; \\ \psi_d &\overset{ind}{\sim} \text{Gamma}(0.5,1),\ d=1,\dots,D; \\ \xi &\sim \text{Half-}\mathcal{N}(0,1). \end{aligned} \tag{11}$$

It can be noted that ξ is a global scale hyperparameter, whereas the independent ψ_d are local scales. The latter ones have Gamma priors with shape parameter 0.5, such value is associated with a more peaked distribution with respect to the Bayesian lasso, encouraging a stronger shrinkage towards 0.

4.1.2. *Posterior inference.* Markov Chain Monte Carlo (MCMC) techniques are particularly suitable for posterior exploration. Specifically, we carry out the fitting by implementing the no-U-turn sampler, an adaptive variant of Hamiltonian Monte Carlo (HMC) algorithm via `Stan` language (Carpenter et al., 2017). We performed estimation by using 4 chains, each with 2,000 iterations, discarding the first 1,000 as warm-up.

Within the Hierarchical Bayes (HB) framework, we assume a quadratic loss and define its posterior expectation as point predictor of θ_d, namely

$$\hat{\theta}_d^{HB} = \mathbb{E}[\theta_d|\text{data}] \quad \forall d, \tag{12}$$

hereafter named model-based estimator. The posterior standard deviation of the target parameter is used to describe its uncertainty.

Users require small area estimates to be robust with respect to model failures. The predictor associated with the popular Fay-Herriot model enjoys an important property in this sense, known as design consistency. Intuitively, it is about the convergence of the model-based predictor to the direct estimator when the area-specific sample size grows large (for a formal definition see Fuller, 2011, p. 41). It can be shown that, when adopting our extended Beta model, $\hat{\theta}_d^{HB}$ is also design-consistent; specifically, conditioning on higher level parameters, we have that $\hat{\theta}_d^{HB} \xrightarrow{p} \hat{\bar{Y}}_d$. A proof of this statement can be found in the supplementary material (De Nicolò, Fabrizi and Gardini, 2024a). In practice, this implies that the difference between the (reliable) direct estimate and the model-based one is negligible in areas with large sample sizes.

4.1.3. *Prediction of out-of-sample areas.* Under the Extended Beta model, we propose in Section 4.1, for the areas that are not included in the sample, the prediction is carried out considering the functional:

$$\theta_d^{OOS} = \mu_d = \text{logit}^{-1}\left(\mathbf{x}_d^T\boldsymbol{\beta} + v_d\right).$$

To obtain a draw from the posterior θ_d^{OOS}, we need one from the distribution $\boldsymbol{\beta}|$data along with one from $v_d|$data. As v_d constitutes a random effect from an unobserved area. Having the b-th Monte Carlo replicate from the posterior distribution $\xi|$data, i.e. $\tilde{\xi}^{(b)}$ we obtain a draw $\tilde{v}_d^{(b)}$ exploiting its hierarchical definition (11):

1. Generate $\tilde{\psi}_d^{(b)}$ from the prior: $\psi_d \sim \text{Gamma}(0.5, 1)$;
2. Generate $\tilde{v}_d^{(b)}$ from $v_d|\tilde{\psi}_d^{(b)}, \tilde{\xi}^{(b)} \sim \mathcal{N}\left(0, \tilde{\psi}_d^{(b)}/\tilde{\xi}^{(b)}\right)$.

4.2. *The Arc-sine model.* For comparison purposes, we consider the Fay-Herriot model with arc-sine square root transformation as an alternative model, commonly used for small area estimation of ratios and proportions. This model is adopted in the context of poverty mapping by Casas-Cordero Valencia, Encina and Lahiri (2016) and Schmid et al. (2017) among others. Frequentist prediction can be implemented in the `emdi` R package (Kreutzmann et al., 2019). Bayesian inference for this model is discussed by Raghunathan et al. (2007). By using the previous notation, the model can be outlined as follows:

$$sin^{-1}\left(\hat{\bar{Y}}_d^{\frac{1}{2}}\right)|\boldsymbol{\beta}, v_d \overset{ind}{\sim} \mathcal{N}(\eta_d, S_d^2),$$

$$\eta_d = \mathbf{x}_d^T\boldsymbol{\beta} + v_d, \quad d = 1, \ldots, D;$$

with S_d^2 being a variance parameter generally assumed to be known. This transformation has a twofold motivation: in the first place, it guarantees that the back-transformed predictor lies in the appropriate proportion range $0 \leq \mathbb{E}[\sin^2(\eta_d|\text{data})] \leq 1$, once the domain of the linear predictor is truncated to the interval $\eta_d \in [0; \pi/2]$. Moreover, it has also the advantage of variance stabilization: the sampling variances for the inverse sine transformed can be approximated by a parameter-free function of the (equivalent) sample size retrieved using the Delta method (Efron and Morris, 1975):

$$S_d^2 \cong \frac{\hat{\bar{Y}}_d(1-\hat{\bar{Y}}_d)}{4\tilde{n}_d\hat{\bar{Y}}_d(1-\hat{\bar{Y}}_d)} = \frac{1}{4\tilde{n}_d}; \tag{13}$$

where the last equality holds only when $\hat{\bar{Y}}_d \in (0; 1)$. We propose an HB approach for estimating the Arc-sine model as in Raghunathan et al. (2007), but with a different prior specification for the unknown parameters to parallel that defined in Subsection 4.1.1. The regularized horseshoe prior for $\boldsymbol{\beta}$ in (9) has been considered with the sole difference of replacing the pseudo variance in (10) with

$$\tilde{\sigma}^2 = \frac{\sum_{d \in D_s}(z_d - \bar{z})^2}{D_s - 1},$$

where $z_d = \sin^{-1}\left(\hat{\bar{Y}}_d^{\frac{1}{2}}\right)$. While the global-local shrinkage prior for v_d has been defined exactly in the same way as in (11). Posterior inference on the target parameter is based on back-transformation. Therefore the HB estimator on the original scale is a result of a proper back-transformation as $\hat{\theta}_d^{HB} = \mathbb{E}[\sin^2(\eta_d|\text{data})]$. The transformation, applied directly on posterior draws, avoids bias issues related to the back-transformation that are common in the frequentist framework (Sugasawa and Kubokawa, 2017). The model estimation has been carried out in line with Section 4.1.2, while estimates for out-of-sample areas consider the functional:

$$\theta_d^{OOS} = \sin^2\left(\mathbf{x}_d^T\boldsymbol{\beta} + v_d\right),$$

with draws from the posterior obtained following the steps defined in Section 4.1.3.

4.3. *The Binomial model.* We introduce a Binomial model as another benchmark. This implies a slight change of perspective since it does not model the survey proportion $\hat{\bar{Y}}_d$, but rather the sample count of poor people in an area. This strategy is discussed by Chen, Wakefield and Lumely (2014) and Benedetti, Berrocal and Little (2022) is adopted in a poverty mapping framework by Franco and Bell (2015). To take into account the complexity of the sampling design, such models use as response the effective number of cases (i.e. poor people) defined as $T_d = [\tilde{n}_d \cdot \hat{\bar{Y}}_d]$, where square brackets denote rounding to the nearest integer. The rounded effective sample size $[\tilde{n}_d]$ is adopted as number of trials. This results in the hierarchical model:

$$T_d|\theta_d \overset{ind}{\sim} \text{Binom}\left(\theta_d, [\tilde{n}_d]\right), \tag{14}$$

$$\text{logit}\left(\theta_d\right) = \mathbf{x}_d^T\boldsymbol{\beta} + v_d, \quad d = 1, \ldots, D. \tag{15}$$

It follows that the target parameter is $\theta_d = \mathbb{E}[T_d|\theta_d]/[\tilde{n}_d]$, provided that the whole procedure overlooks the impact of rounding $\tilde{n}_d \cdot \hat{\bar{Y}}_d$ and $\tilde{n}_d$. Similarly, $\mathbb{V}[T_d|\theta_d]/[\tilde{n}_d] = \theta_d(1-\theta_d)[\tilde{n}_d]^{-1}$ mimics the variance of the frequency under a Binomial random variable, that assumes independence among trials. Such expression is also coherent with the sampling variance under the Beta model. The Binomial working likelihood naturally accommodates $\hat{\bar{Y}}_d$ equal to either 0 or 1, implying $T_d = 0$ or $T_d = [\tilde{n}_d]$. Due to the rounding process, we point out that it is possible to observe $T_d = 0$ when $\hat{\bar{Y}}_d \approx 0$ and/or $\tilde{n}_d \approx 0$, whereas $T_d = [\tilde{n}_d]$ when $\hat{\bar{Y}}_d \approx 1$. Under the model, it is straightforward to get $\mathbb{P}[T_d = 0] = (1-\theta_d)^{[\tilde{n}_d]}$ and $\mathbb{P}[T_d = [\tilde{n}_d]] = \theta_d^{[\tilde{n}_d]}$.

The model is fitted under an HB approach with prior specification coherent with the one exposed in Subsection 4.1.1. To calibrate the horseshoe prior (9) on the coefficients, we replace the expression (10) for the pseudo-variance with

$$\tilde{\sigma}^2 = \frac{1}{\bar{\mu}(1-\bar{\mu})}, \tag{16}$$

in line with Piironen and Vehtari (2017). Lastly, inference for out-of-sample areas follows Subsection 4.1.3 with $\theta_d^{OOS} = \text{logit}^{-1}\left(\mathbf{x}_d^T\boldsymbol{\beta} + v_d\right)$.

5. Design-based simulation. In this section, we introduce a design-based simulation to assess the frequentist properties of model-based estimates obtained under Extended Beta (EB), Arc-sine (AS) and Binomial models. We also introduce in the comparison the model by Fabrizi, Ferrante and Trivisano (2016) (FFT model), in order to measure the impact of relaxing the independence assumption. The simulation study is design-based to avoid data generation under specific model assumptions; we rather try to reproduce a framework that is as close as possible to real poverty data.

We assume the DHS sample as a synthetic population and the 64 zilas as domains. Then $B = 1000$ samples are drawn from the synthetic population by mimicking the DHS design, including stratification and multi-stage selection. We draw samples made of 114 clusters stratifying by zilas in order to control for the domain-specific sample sizes; 10 zilas with 3 or fewer clusters in the synthetic population are considered out-of-sample areas. From each cluster, 25% of households are randomly selected. This implies samples of different sizes at each iteration: on average, 5.84% of the population is sampled at each iteration, with domain sample sizes ranging from 27.94 to 260.09, with a mean of 73.22. For each sample, direct estimates are computed and used as input for the four small area models involved in the simulation study. They provide the following model-based estimators:

1. The empirical best linear predictor (EBLUP) under the Arc-sine model (EBLUP-AS) provided by the package `emdi`;

	In-Sample areas					
	Direct Est.	EBLUP-AS	HB-AS	HB-Bin	HB-EB	HB-FFT
RMSE	0.142	0.086	0.078	0.075	0.071	0.076
BIAS	0.000	-0.010	-0.021	0.001	-0.009	0.004
90% Cov.	-	-	0.893	0.920	0.933	0.911
	Out-of-Sample areas					
	Direct Est.	EBLUP-AS	HB-AS	HB-Bin	HB-EB	HB-FFT
RMSE	-	0.154	0.110	0.123	0.118	0.123
BIAS	-	0.058	-0.043	-0.006	0.030	0.028
90% Cov.	-	-	0.978	0.937	0.930	0.933

TABLE 2

Median Bias, RMSE and frequentist coverage for the different estimation methods considered, distinguishing between sampled areas and out-of-sample.

2. The HB estimator under the Arc-sine model (HB-AS);
3. The HB estimator under the Binomial model (HB-Bin);
4. The HB estimator under the Extended Beta model (HB-EB);
5. The HB estimator under the FFT model (HB-FFT).

We exploit the Monte Carlo variances of estimators to compute the area-specific effective sample sizes and the spatial covariates at zila level are obtained following the same methodology of Section 2.2. The whole set of available covariates is provided as input to models HB-AS, HB-Bin, HB-EB and HB-FFT, whereas a preliminary model selection step is required for EBLUP-AS, in order to obtain an optimal subset. The frequentist procedure would not simply work with the large number of covariates we computed. Specifically, we carry out the selection by fitting the model with the synthetic population data and using AIC as the selection criterion. Clearly, in this way, the EBLUP-AS strategy relies on different modeling conditions and it is not directly comparable to the Bayesian procedures, automatically incorporating the model selection step. The uncertainty involved in the regressors selection step is overlooked.

Let us denote with $\hat{\theta}_{db}$ the model-based estimate for domain d at iteration b with population value θ_d; we consider bias, root mean squared error (RMSE) and frequentist coverage of the 90% credible intervals to compare estimators performances. Such quantities are defined as:

$$\text{Bias}\left(\hat{\theta}_d\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{db} - \theta_d\right), \quad \text{MSE}\left(\hat{\theta}_d\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{db} - \theta_d\right)^2,$$

$$\text{Coverage}_{90}(\hat{\theta}_d) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\left\{\theta_d \in [Q_{0.05}(\theta_{db}|\text{data}), Q_{0.95}(\theta_{db}|\text{data})]\right\},$$

where $Q_\alpha(\theta_{db}|\text{data})$ denotes the posterior quantile of order α of θ_{db}.

In Table 2, the medians of area-specific biases and RMSEs are reported, including also the performances of the direct estimator as a benchmark. By focusing on in-sample areas, we remark that the considered small-area models behave rather similarly. As expected, the direct estimator is unbiased and, among model-based estimators, a slight negative median bias is registered for the HB-AS model. Focusing on the RMSE, we can first note the remarkable decrease yielded by the use of small area models with respect to direct estimation. In terms of median, the HB-EB model shows a lower RMSE compared to other models; specifically, the HB-EB model has a smaller RMSE with respect to other proposals in approximately 7 out of 10 areas (i.e. 74.1% versus HB-FFT, 68.5% versus HB-AS, and 66.7% versus HB-Binom).

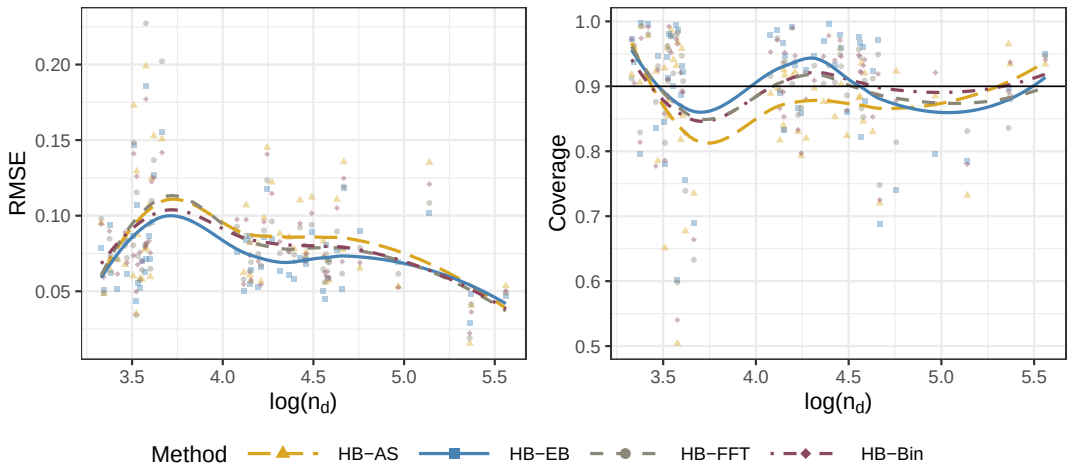

FIG 2. *Behaviour of RMSE and frequentist coverage with respect to the area sample size* n_d.

The left plot of Figure 2 shows the behavior of RMSEs with respect to the log of the average area sample size: the LOESS smoothing line related to the HB-EB model is systematically below the ones related to the HB-Bin, HB-AS and HB-FFT models. Table 2 also reports the results concerning out-of-sample areas. Comparable results are obtained, we note that the EBLUP-AS model shows a higher RMSE and positive bias, making it less reliable in case of out-of-sample prediction.

Focusing on the frequentist coverage for the 90% credible interval, we note how the median coverage, reported in Table 2, is satisfactory for all the Bayesian methods as they reach the nominal level, with a slight tendency to over-coverage of HB-EB. For details about the area-specific coverages with respect to the sample size, see the right plot of Figure 2. We note that the coverage is occasionally very low, especially for areas with tiny samples; this is due to the strong synthetic component of the predictors and the somewhat deviant behavior of these areas. Similar coverage values are obtained for the out-of-sample areas which represent a valuable result, confirming that the procedure described in Section 4.1.3 propagates uncertainty successfully. Lastly, the HB-AS method shows a marked tendency to overshoot the *nominal* coverage level.

A possible limitation of our simulation is that being fully based on DHS data as those of our application, comparisons should not be as general or conclusive. Nonetheless, the HB-EB model seems to work slightly better in this context. The first motivation is that the Beta likelihood accommodates the potential skewness of sampling distribution better than a Gaussian approximation of the transformation. A second possible motivation is that the Arc-sine models use a variance approximation on the transformed scale which is known to fail when true probabilities are very close to 0 (Efron and Morris, 1975), as it is often the case in our setting.

6. Application on Bangladesh DHS data. In this section, we map poverty in the Bangladeshi upazilas by integrating the DHS data and remote sensing covariates described in Section 2. We remark that the DHS dataset is composed of 365 in-sample areas with direct estimates ranging from 0 to 0.96 (median: 0.16, 66 zero values), while 179 areas are out-of-sample. The section is divided into two parts: in Subsection 6.1, we assess why the assumptions and approximations at the base of the Arc-sine model are not suitable for the analyzed data. For this reason, we do not present the results pertaining to the Arc-sine model in Subsection 6.2, where we illustrate estimates and diagnostics for models based on the Beta

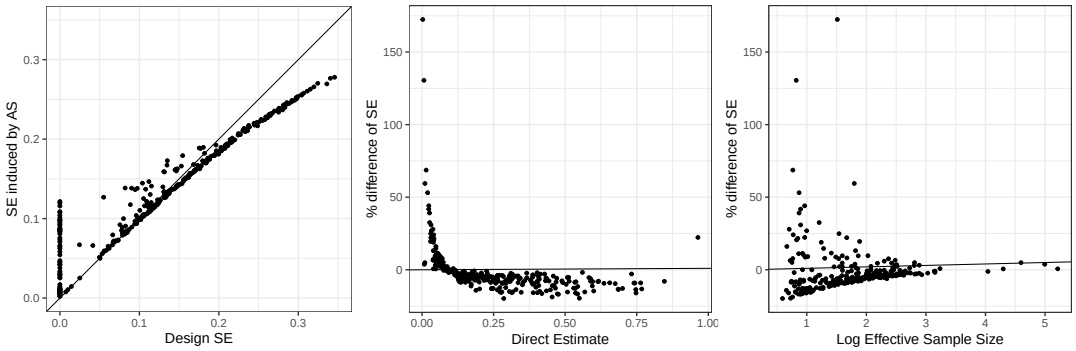

FIG 3. *Design standard error of $\hat{\bar{Y}}_d$ versus standard errors implicitly assumed by the AS model (left panel). Their percentage differences are compared to direct estimates (center) and $\log \tilde{n}_d$ (right panel).*

distribution (EB and FFT) and the Binomial model. Lastly, the Subsection 6.3 contains some comments and remarks on poverty mapping in Bangladesh.

6.1. *Why we need to go beyond the Arc-sine model.* Our data are characterized by the presence of many zeros among direct estimates. While the extended Beta model is built with the aim of separately modeling such values, the Arc-sine model is usually employed treating them as any value in the domain $[0; 1]$. This could introduce some counter-intuitive consequences, in particular when determining the variances of transformed direct estimates.

Being the response a proportion, associating a non-null variance to 0 values is not in accordance with the theory of binomial processes. Nonetheless, this is what is done when the approximation to the variance (13), i.e. $(4\tilde{n}_d)^{-1}$, is applied to compute the standard error of transformed direct estimates, including those equal to 0. Actually, the approximation can be applied only when $\hat{\bar{Y}}_d \in (0; 1)$. In addition, its accuracy decays near 0 and 1 values and especially so when effective sample sizes are small.

To check the implication of the variance approximation (13) in our analysis, we compute by simulation the standard errors associated with the (untransformed) $\hat{\bar{Y}}_d$ induced by the Arc-sine model. We proceed as follows: we draw Monte Carlo samples from a $\mathcal{N}\left(\sin^{-1}\left(\hat{\bar{Y}}_d^{\frac{1}{2}}\right), (4\tilde{n}_d)^{-1}\right)$, then we apply the back-transformation and compute standard errors for $\hat{\bar{Y}}_d$ from the back-transformed samples.

In Figure 3, such quantities are compared with $\hat{\mathbb{SE}}_{cs}[\hat{\bar{Y}}_d]$, computed as in (3). The approximation leads to a huge overestimation of standard error (up to 170%) for estimates close to the boundaries and an underestimation (up to -20%) otherwise. The amount of such discrepancies increases with a low effective sample size, as clear from the right-hand side plot in Figure 3.

As many zeros or close-to-zero values are present, the approximation seems not adequate for our data, with a relevant impact on the model-based estimates with respect to those relying on more accurate design standard errors. The Extended Beta model avoids this problem since it operates in the original scale without any approximation, dealing with 0 values separately. Eventually, we also note that the Beta-based models offer the additional pro of favoring the interpretability of regression coefficients: as a logit link is used, the exponentiated regression slopes can be read in terms of probability odds (Warton and Hui, 2011). For these reasons, we decide to rule out the Arc-sine model from the following analysis.

6.2. *Comparing Beta-based and Binomial models.* In this subsection, we discuss the estimation and results interpretation of Binomial, EB and FFT models. As for specification, the

	EB Model				FFT Model			
	Post. Mean	Post. SD	$Q_{0.025}$	$Q_{0.975}$	Post. Mean	Post. SD	$Q_{0.025}$	$Q_{0.975}$
ξ	0.13	0.10	0.01	0.36	0.99	0.13	0.74	1.27
λ	0.80	0.02	0.75	0.84	-	-	-	-
LOOIC (SE)	-118.9 (34.1)				-14.7 (44.2)			

TABLE 3

Posterior summaries of parameters ξ and λ and LOOIC.

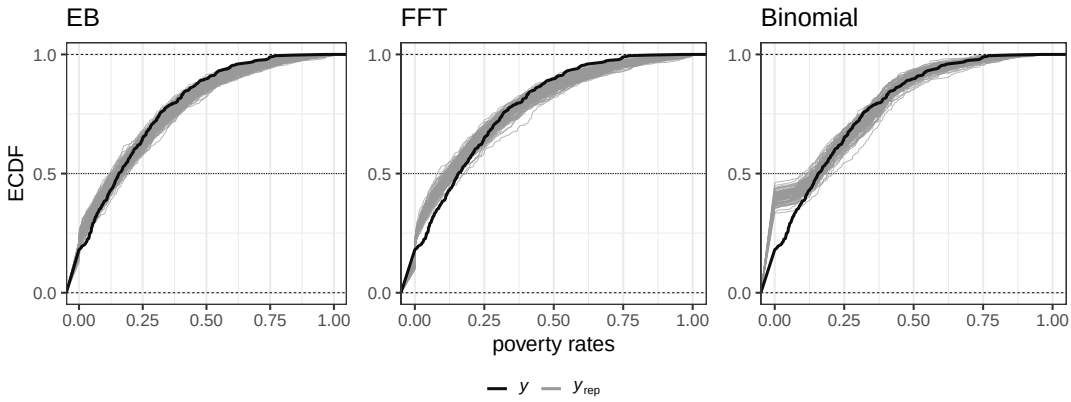

FIG 4. *ECDF of the posterior predictive distributions under models EB, FFT and Binomial (in gray) compared to the ECDF of the set of direct estimates (in black).*

horseshoe priors described in Section 4.1.1 need to be completed with additional hyperparameters settings. Specifically, the expected number of relevant coefficients has been set to $p_0 = 10$, according to the results of a preliminary regressors selection exercise. Moreover, the data pseudo-variances obtained by applying (10) and (16) result to be $\tilde{\sigma} = 1.23$ for the Beta model and $\tilde{\sigma} = 2.49$ for the Binomial one.

Regarding model comparison, our attention is on posterior predictive checks, as discussed in Gabry et al. (2019). In Figure 4, we conduct a comparative analysis of empirical cumulative distribution functions (ECDFs). Specifically, we contrast the ECDF of direct estimates (depicted in black) with those derived from samples generated from the posterior predictive distribution of various models (depicted in grey). This examination exposes deficiencies in the FFT and Binomial models when handling the probabilities of observing zero values. Issues with the FFT model stem from its independence assumption, which results in a slight underestimation of the probability of observing direct estimates equal to zero. Conversely, the Binomial model exhibits evident miscalibration near the zero value, consistently overshooting the probability of observing proportions equal to zero. This behavior can be attributed to two potential factors. Firstly, it may arise from the rounding step in the construction of the response variable, wherein $\{\hat{\bar{Y}}_d = 0\} \subset \{T_d = 0\}$ inflates the probability of zero counts. Secondly, the rigidity in modeling $\mathbb{P}[T_d = 0]$ as $(1 - \theta_d)^{[\tilde{n}_d]}$: since it relies solely on $[\tilde{n}_d]$, it does not distinguish between intra-cluster correlation and sample sizes. Turning our attention to small area diagnostics, we observe that model-based estimators exhibit significantly lower standard deviations compared to direct estimators (Figure 5). Specifically, the estimations provided by the EB model demonstrate greater reliability in contrast to the FFT and Binomial models.

A further tool of model comparison is the leave-one-out information criterion (LOOIC, Vehtari, Gelman and Gabry, 2017). We remark that this diagnostic can be only exploited to

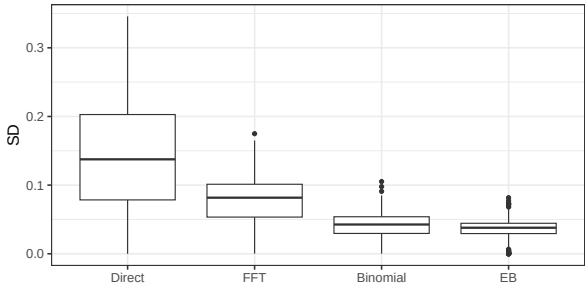

FIG 5. *Distributions of the posterior standard deviations of HB estimates under models EB, FFT and Binomial compared to the standard errors of Direct estimates.*

Covariate	Transformation	$\mathbb{E}\left[\beta_j \vert \text{data}\right]$	Odds Ratio	Importance
VIIRS	Square root	-1.23	0.29	1.00
Woman/Child	Identity	-0.32	0.73	0.98
Distance from woody areas	Log	0.05	1.05	0.80
Time to access the nearest city	Square root	0.80	2.22	0.80
Slope	Inverse	0.05	1.05	0.80
Distance from Coastline	Identity	0.04	1.04	0.71
Distance from vegetation areas	Square root	-0.02	0.98	0.71
Male/Female	Identity	0.06	1.06	0.71

TABLE 4
Posterior summaries of the regression coefficients β_j

compare the Beta-based models, being the response used in the Binomial model substantially different. In detail, Table 3 shows lower values of LOOIC for the EB model, being preferable as compared to FFT model. This points out that the introduction of the correlation parameter λ allows us to better model the poverty rates near zero, as can be seen from the other posterior summaries in Table 3. We observe that such a parameter is well identified by the data, with a posterior mean equal to 0.80. Note that the biggest $\mathbb{E}[\mu_d|\text{data}]$ reaches 0.58, being $\mathbb{E}[\mu_d|\text{data}] < \mathbb{E}[\lambda|\text{data}] < 1$, $\forall d$. This confirms the presence of a strong positive correlation among sampled households as already observed by intra-cluster correlation estimates of Section 3.1. Lastly, we highlight that the misspecification in the modeling of censored values probabilities induces an increase in random effect variability. Indeed, the global scale parameter ξ of the variance gamma prior is estimated ten times larger in the FFT model.

6.3. *Poverty mapping in Bangladesh.* After showing that the proposed EB model is suitable to fit the analysed data, we present the obtained results of the poverty mapping analysis carried out on Bangladeshi data. Table 4 reports the posterior summaries of regression coefficients for the most important covariates, i.e. those with high *Importance*, that we define as

$$\max(\mathbb{P}[\beta_j < 0|\text{data}], \mathbb{P}[\beta_j > 0|\text{data}]).$$

The VIIRS covariate has the greatest importance with an expected negative sign, as the most enlightened areas during nighttime are characterized by smaller poverty rates. Among the demographic covariates, the most relevant one is the woman/child dependency ratios, being inversely proportional to the probability of being poor, as expected. Among other covariates in the top list of importance, we note *Time to access the nearest city*. As already discussed in the literature (Iimi et al., 2016; Islam, Sayeed and Hossain, 2017), remoteness and exclusion

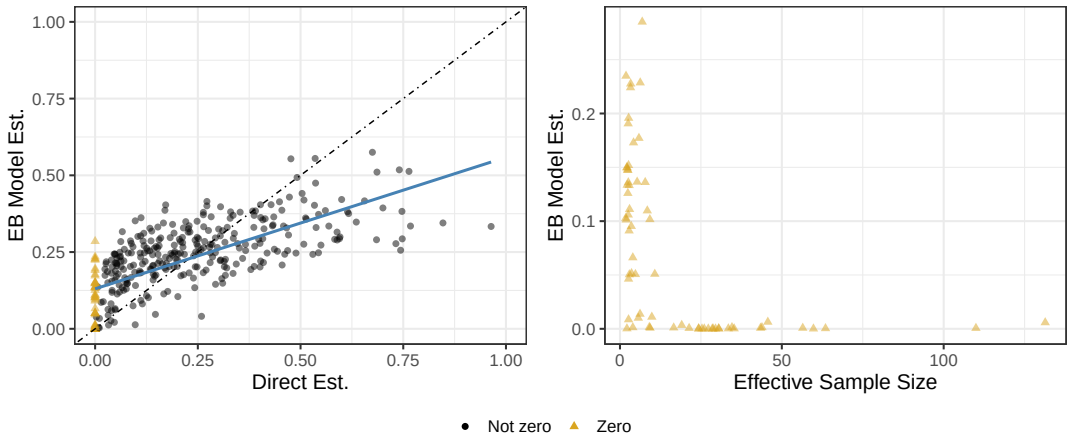

FIG 6. *Left panel: EB model estimates versus direct estimates (bisector as dash-dotted line, linear regression line as solid line). Right panel: model-based estimates versus effective sample sizes for areas with 0-valued direct estimates.*

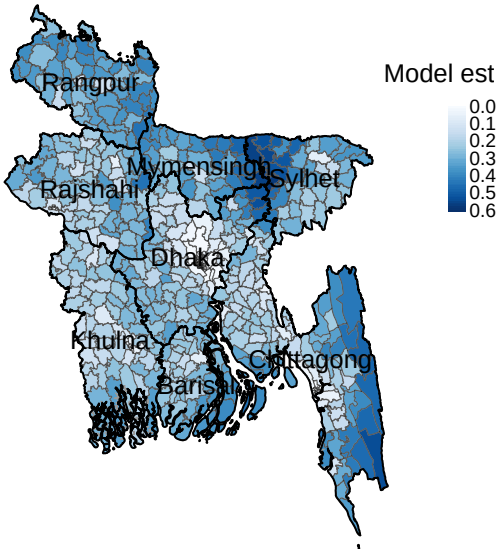

FIG 7. *Map with model-based estimates of poverty rates in Bangladeshi upazilas.*

from the national labor and goods market represent one of the main drivers of poverty at the community level in Bangladesh.

The amount of shrinkage induced by the model is described in the left panel of Figure 6, i.e. direct estimates versus EB-based ones; it is strong as expected, given the low precision of direct estimates. Zero estimates (highlighted in golden) are clearly shrunk towards the center of the distribution. The right panel of Figure 6 displays how zero-valued direct estimates are spread by the model with respect to the effective sample sizes. Note that the impact model has on zero estimates is mainly restricted to extremely small sizes. We remark on the presence of a subset of upazilas with very small poverty rates, which are mostly located in urban districts. The urban-rural divide is still an important catalyst for poverty differences (Khudri et al., 2013; Islam, Sayeed and Hossain, 2017) as far as wealth indicators are concerned.

A map of model-based poverty estimates at the upazila level can be found in Figure 7. As compared to the direct estimates mapped in Figure 1, we see how model predictions fill the many grey areas (out-of-sample), especially in peripheral regions. Large disparities among regions are clearly noticeable: for instance, the metropolitan regions of Dhaka and Chattogram retain the lowest poverty levels, while those far from cities, coastlines and roads have the highest. The poverty patterns in Figure 7 are coherent with the ones highlighted by Kam et al. (2005) and Imam et al. (2019). The domains with high poverty incidence overlap with areas ecologically poor for food production: the depression area in the north-east, called Haor (Sylhet basin lowlands) and some areas at the edge of major rivers, such as the Jamuna river, both particularly exposed to climate change effects and floodings (Haque and Jahan, 2015); the drought-prone area of Rangpur division in the north-west and the remote area on the Chittagong Hill Tract (south-east). The western part, around southern Rajshahi and Khulna division, has lower poverty levels since, even if drought-prone, it presents good irrigation coverage (Kam et al., 2005). We have no clear evidence of the East-West divide (World Bank, 2008), in line with recent literature highlighting its decreasing relevance (Rahman et al., 2017).

The model-based estimates in Figure 7 appear to be spatially smooth and clustered. To investigate a possible residual spatial trend, we perform the Moran's and Geary's tests for spatial autocorrelation on the residuals of the EB model. Both tests do not reject the null hypothesis of zero autocorrelation (p-value equal to 0.41 and 0.33, respectively). In addition, we extend the EB model with a spatially structured random effect in the linear predictor, having an Intrinsic Conditional Autoregressive prior in line with Porter et al. (2014). Results highlight the non-relevance of the spatial term in the analysis: the LOOIC is -119.0, i.e. almost equal to the one related to the EB model in Table 3; without systematic differences in area estimates. A reason may be that the spatial correlation of direct estimators conditionally on covariates is negligible as the set of remote sensing covariates is able to largely explain it. For further details, see Section S3 of the Supplementary Material (De Nicolò, Fabrizi and Gardini, 2024a). Figure 8 displays the map of the posterior standard deviation on the right panel compared with the standard error of direct estimates on the left one. Note that the posterior standard error is not only lighter but also more homogeneous since small area predictors are dominated by the synthetic part. Being reliable, model-based estimates can be released and employed for further analysis.

7. Conclusions and directions for further research. The applied problem of mapping poverty in Bangladesh by integrating a survey sample and remote sensing data drove us to set up a novel hierarchical Bayesian model based on the Beta likelihood. We did this in the line of small area literature relying on area-level models. Our purpose was to provide a more general tool for poverty mapping in developing countries with respect to existing alternatives, ensuring a convenient implementation that requires no auxiliary variable selection and minimal intervention in the prior specification. Indeed, the latter aspect has often represented a limit to the widespread use of Bayesian methods among practitioners. As far as computation is concerned, the code implementing the whole procedure is available in the supplementary material (De Nicolò, Fabrizi and Gardini, 2024b), together with a toy dataset to test implementation. Furthermore, our methodology is going to be released in the R `tipsae` package (De Nicolò and Gardini, 2024), complementing the set of tools for Bayesian small area estimation of proportions and indicators in the unit interval.

From a methodological point of view, we specified an Extended Beta mixed regression model. We extended the proposal of Fabrizi, Ferrante and Trivisano (2016) to more effectively handle data features as the presence of estimates equal to either 0 or 1 and a strong intra-cluster correlation of observations. The simulation and application results underline the

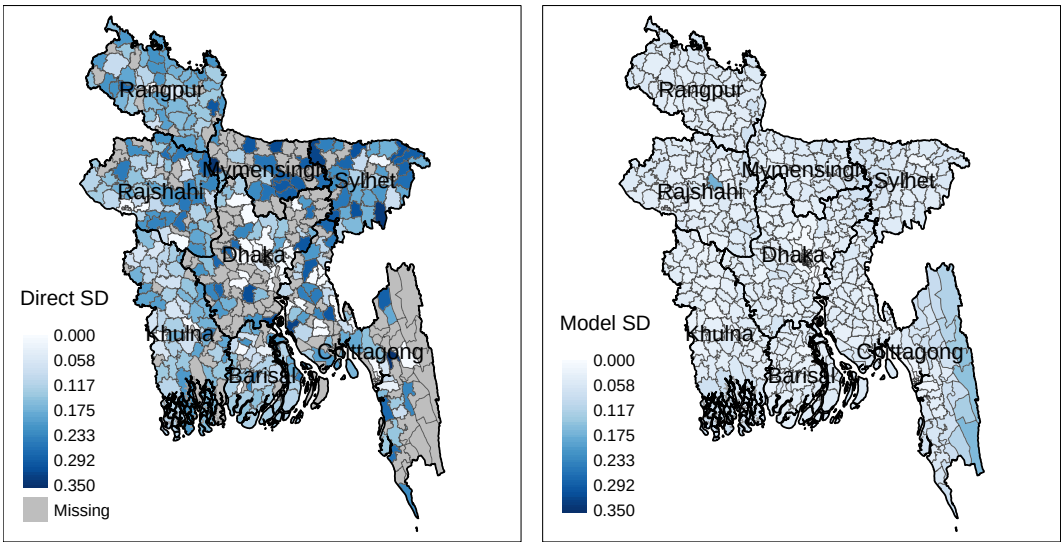

FIG 8. *Standard error of direct estimates and posterior standard deviation of model-based estimates.*

importance of the additional correlation parameter, sensibly improving goodness-of-fit and leading to more precise estimates. Moreover, the explicit probabilistic formulation placed on the occurrence of observing zero/one values makes the EB model more interpretable with respect to other proposals (Warton and Hui, 2011).

An important point that emerges from the results reported in the paper is that some challenges that characterize the analyzed data require ad hoc modeling solutions. In particular, the proposed model outperforms other ready-to-use methodologies discussed in the small area literature (i.e., the Arc-sine FH model, the FFT model, and the Binomial model) in taking into account the strong intra-cluster correlation and the high incidence of direct estimates equal to zero. We do not consider other alternative models, e.g. the Poisson model, for which available proposals from the literature require a tricky extension for our application. For example, Boubeta, Lombardía and Morales (2017) do not consider the complex survey design. Indeed, including the sampling variances under a Poisson model is not trivial and it requires radically different assumptions on the uncertainty evaluation of estimates which are beyond the scope of this paper (e.g. Bradley, Wikle and Holan, 2016), being an interesting topic for future research.

Our research concerning this applied problem is not over. If we consider administrative units larger than upazilas, we expect direct estimates to gain precision and remote sensing predictors to lose predictive power, being averaged on a wider area. This may impact small area results raising up the need to combine different information layers at once.

Acknowledgments. Work supported by the Data and Evidence to End Extreme Poverty (DEEP) research programme. DEEP is a consortium of the Universities of Cornell, Copenhagen, and Southampton led by Oxford Policy Management, in partnership with the World Bank - Development Data Group and funded by the UK Foreign, Commonwealth & Development Office. The work of Silvia De Nicolò was partially supported by the ALMA IDEA 2022 grant (title: "Social exclusion and territorial disparities: poverty and inequality mapping through advanced methods of small area estimation", project J45F21002000001), funded by the European Union - NextGenerationEU and PNRR funds, PE10 project – ONFOODS, "Research and innovation network on food and nutrition. Sustainability, Safety and Security

– Working ON Foods" (code PE0000003, CUP J33C22002860001). The work of Aldo Gardini was partially supported by MUR on funds FSE REACT EU - PON R&I 2014-2020 and PNR (D.M. 737/2021) for the RTDA_GREEN project (title: "Modelli statistici per lo studio della convergenza spaziale verso la transizione verde", J41B21012140007).

SUPPLEMENTARY MATERIAL

Supplementary Material Document

Additional information on Remote Sensing (RS) covariates, used as auxiliary variables in the small area model, are supplied such as the raw features and data sources. Moreover, the proof of the design consistency property related to the Extended Beta model-based estimator is provided and the formalization and estimation of the spatial model are discussed.

R Code

Code to run the `Stan` model with pseudo-data.

REFERENCES

BENEDETTI, M. H., BERROCAL, V. J. and LITTLE, R. J. (2022). Accounting for survey design in Bayesian disaggregation of survey-based areal estimates of proportions: an application to the American Community Survey. *The Annals of Applied Statistics* **16** 2201–2230.

BOUBETA, M., LOMBARDÍA, M. J. and MORALES, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics & Data Analysis* **107** 32–47.

BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American Community Survey. *Journal of the American Statistical Association* **111** 472–487.

BROWN, P. J. and GRIFFIN, J. E. (2010). Inference with Normal-Gamma prior distributions in regression problems. *Bayesian Analysis* **5** 171–188.

BURGERT, C. R., ZACHARY, B. and COLSTON, J. (2013). Incorporating geographic information into demographic and health surveys: a field guide to GPs data collection. *Calverton, Maryland, USA: ICF International*.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software* **76**.

CASAS-CORDERO VALENCIA, C., ENCINA, J. and LAHIRI, P. (2016). Poverty mapping for the Chilean comunas. *Analysis of Poverty Data by Small Area Estimation* 379–404.

CHEN, S. and RUST, K. (2017). An extension of Kish's formula for design effects to two-and three-stage designs with stratification. *Journal of Survey Statistics and Methodology* **5** 111–130.

CHEN, C., WAKEFIELD, J. and LUMELY, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spatial and spatio-temporal epidemiology* **11** 33–43.

CORSI, D. J., NEUMAN, M., FINLAY, J. E. and SUBRAMANIAN, S. (2012). Demographic and health surveys: a profile. *International journal of epidemiology* **41** 1602–1613.

DATTA, G. S., HALL, P. and MANDAL, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association* **106** 362–374.

DE NICOLÒ, S. and GARDINI, A. (2024). The R package tipsae: Tools for mapping proportions and indicators on the unit interval. *Journal of Statistical Software* **108** 1–36.

DE NICOLÒ, S., FABRIZI, E. and GARDINI, A. (2024a). Supplementary material for "Extended Beta models for poverty mapping. An application integrating survey and remote sensing data in Bangladesh".

DE NICOLÒ, S., FABRIZI, E. and GARDINI, A. (2024b). Supplementary R code for "Extended Beta models for poverty mapping. An application integrating survey and remote sensing data in Bangladesh".

DURANTON, G. and VENABLES, A. J. (2021). Place-based policies: principles and developing country applications. In *Handbook of regional science* 1009–1030. Springer.

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70** 311–319.

ENGSTROM, R., HERSH, J. S. and NEWHOUSE, D. L. (2017). Poverty from space: using high-resolution satellite imagery for estimating economic well-being. *World Bank Policy Research Working Paper* **8284**.

FABRIZI, E., FERRANTE, M. and TRIVISANO, C. (2016). Hierarchical Beta regression models for the estimation of poverty and inequality parameters in small areas. *Analysis of Poverty Data by Small Area Methods. John Wiley and Sons* 299–314.

FABRIZI, E., FERRANTE, M. R. and TRIVISANO, C. (2018). Bayesian small area estimation for skewed business survey variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67** 861–879.

FABRIZI, E., FERRANTE, M. R. and TRIVISANO, C. (2020). A functional approach to small area estimation of the Relative Median Poverty Gap. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **183** 1273–1291.

FABRIZI, E., FERRANTE, M. R., PACEI, S. and TRIVISANO, C. (2011). Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Computational Statistics & Data Analysis* **55** 1736–1747.

FAY, R. E. and 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* **74** 269–277.

FERRARI, S. and CRIBARI-NETO, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31** 799–815.

FRANCO, C. and BELL, W. R. (2015). Borrowing information over time in binomial/logit normal models for small area estimation. *Statistics in Transition new series* **4** 563–584.

FULLER, W. A. (2011). *Sampling Statistics*. John Wiley & Sons.

GABLER, S., HÄDER, S. and LAHIRI, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology* **25** 105–106.

GABRY, J., SIMPSON, D., VEHTARI, A., BETANCOURT, M., GELMAN, A. et al. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society Series A* **182** 389–402.

HÁJEK, J. (1971). Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu. *Foundations of Statistical Inference* 326.

HAQUE, A. and JAHAN, S. (2015). Impact of flood disasters in Bangladesh: A multi-sector regional analysis. *International Journal of Disaster Risk Reduction* **13** 266–275.

HAY, S. I. and SNOW, R. W. (2006). The Malaria Atlas Project: developing global maps of malaria risk. *PLoS medicine* **3** e473.

IIMI, A., AHMED, F., ANDERSON, E. C., DIEHL, A. S., MAIYO, L., PERALTA-QUIRÓS, T. and RAO, K. (2016). New rural access index: main determinants and correlation to poverty. *World Bank Policy Research Working Paper* **7876**.

IMAM, M. F., ISLAM, M. A., ALAM, M. A., HOSSAIN, M. J. and DAS, S. (2019). Small Area Estimation of Poverty in Rural Bangladesh. *The Bangladesh Journal of Agricultural Economics* **40** 1–16.

ISLAM, D., SAYEED, J. and HOSSAIN, N. (2017). On determinants of poverty and inequality in Bangladesh. *Journal of Poverty* **21** 352–371.

JANICKI, R. (2020). Properties of the Beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Communications in Statistics-Theory and Methods* **49** 2264–2284.

KALTON, G. (1979). Ultimate cluster sampling. *Journal of the Royal Statistical Society: Series A (General)* **142** 210–222.

KAM, S.-P., HOSSAIN, M., BOSE, M. L. and VILLANO, L. S. (2005). Spatial patterns of rural poverty and their relationship with welfare-influencing factors in Bangladesh. *Food Policy* **30** 551–567.

KHUDRI, M. M., CHOWDHURY, F. et al. (2013). Evaluation of socio-economic status of households and identifying key determinants of poverty in Bangladesh. *European Journal of Social Sciences* **37** 377–387.

KISH, L. (1987). Weighting in Deft2. *The Survey Statistician* **17** 26–30.

KLOTZ, J. (1973). Statistical inference in Bernoulli trials with dependence. *The Annals of statistics* 373–379.

KREUTZMANN, A.-K., PANNIER, S., ROJAS-PERILLA, N., SCHMID, T., TEMPL, M. and TZAVIDIS, N. (2019). The R Package `emdi` for Estimating and Mapping Regionally Disaggregated Indicators. *Journal of Statistical Software* **91** 1–33. https://doi.org/10.18637/jss.v091.i07

LIU, B., LAHIRI, P. and KALTON, G. (2007). Hierarchical Bayes modeling of survey-weighted small area proportions. In *Proceedings of the American Statistical Association, Survey Research Section* 3181–3186.

LYNN, P., HÄDER, S. and GABLER, S. (2006). Design effects for multiple design samples. *Survey Methodology* **32** 115–120.

MARHUENDA, Y., MOLINA, I. and MORALES, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics & Data Analysis* **58** 308–325.

MASAKI, T., NEWHOUSE, D., SILWAL, A. R., BEDADA, A. and ENGSTROM, R. (2020). Small area estimation of non-monetary poverty with geospatial data.

MOLINA, I., NANDRAM, B. and RAO, J. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics* **8** 852–885.

NIPORT AND MITRA AND ASSOCIATES AND ICF INTERNATIONAL (2016). Bangladesh Demographic and Health Survey 2014 Technical Report, National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International, Dhaka, Bangladesh, and Rockville, Maryland, USA.

O'DONNELL, M. S. and IGNIZIO, D. A. (2012). Bioclimatic predictors for supporting ecological applications in the conterminous United States. *US geological survey data series* **691** 4–9.

PEZZULO, C., HORNBY, G. M., SORICHETTA, A., GAUGHAN, A. E., LINARD, C., BIRD, T. J., KERR, D., LLOYD, C. T. and TATEM, A. J. (2017). Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. *Scientific Data* **4** 1–15.

PIIRONEN, J. and VEHTARI, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11** 5018 – 5051. https://doi.org/10.1214/17-EJS1337SI

POIRIER, M. J., GRÉPIN, K. A. and GRIGNON, M. (2020). Approaches and alternatives to the wealth index to measure socioeconomic status using survey data: a critical interpretive synthesis. *Social Indicators Research* **148** 1–46.

PORTER, A. T., HOLAN, S. H., WIKLE, C. K. and CRESSIE, N. (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. *Spatial Statistics* **10** 27–42.

RAGHUNATHAN, T. E., XIE, D., SCHENKER, N., PARSONS, V. L., DAVIS, W. W., DODD, K. W. and FEUER, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association* **102** 474–486.

RAHMAN, M. et al. (2017). Role of agriculture in Bangladesh economy: uncovering the problems and challenges. *International Journal of Business and Management Invention* **6**.

RAO, J. N. and MOLINA, I. (2015). *Small area estimation*. John Wiley & Sons.

RIDOUT, M. S., DEMETRIO, C. G. and FIRTH, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55** 137–148.

SCHMID, T., BRUCKSCHEN, F., SALVATI, N. and ZBIRANSKI, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180** 1163–1190.

STEELE, J. E., SUNDSØY, P. R., PEZZULO, C., ALEGANA, V. A., BIRD, T. J., BLUMENSTOCK, J., BJELLAND, J., ENGØ-MONSEN, K., DE MONTJOYE, Y.-A., IQBAL, A. M. et al. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* **14** 20160690.

SUGASAWA, S. and KUBOKAWA, T. (2017). Transforming response values in small area prediction. *Computational Statistics & Data Analysis* **114** 47–60.

TANG, X., GHOSH, M., HA, N. S. and SEDRANSK, J. (2018). Modeling random effects using global–local shrinkage priors in small area estimation. *Journal of the American Statistical Association* **113** 1476–1489.

TATEM, A. J. (2017). WorldPop, open data for spatial demography. *Scientific Data* **4** 1–4.

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27** 1413–1432.

WARTON, D. I. and HUI, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology* **92** 3–10.

WIECZOREK, J. and HAWALA, S. (2011). A Bayesian zero-one inflated beta model for estimating poverty in US counties. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association*.

WORLD BANK (2008). *Poverty Assessment for Bangladesh: Creating Opportunities and Bridging the East-West Divide*. World Bank.

XIE, D., RAGHUNATHAN, T. E. and LEPKOWSKI, J. M. (2007). Estimation of the proportion of overweight individuals in small areas—a robust extension of the Fay–Herriot model. *Statistics in Medicine* **26** 2699–2715.

ZHAO, X., YU, B., LIU, Y., CHEN, Z., LI, Q., WANG, C. and WU, J. (2019). Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh. *Remote Sensing* **11** 375.

ZHOU, Y., MA, T., ZHOU, C. and XU, T. (2015). Nighttime light derived assessment of regional inequality of socioeconomic development in China. *Remote Sensing* **7** 1242–1262.