Silvia De Nicolò, Enrico Fabrizi, Aldo Gardini

Extended Beta Models for Poverty Mapping. An Application Integrating Survey and Remote Sensing Data in Bangladesh

Quaderni di Dipartimento

Serie Ricerche 2022, n. 3
ISSN 1973-9346

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Scienze Statistiche "Paolo Fortunati"

Extended Beta Models for Poverty Mapping. An Application Integrating Survey and Remote Sensing Data in Bangladesh

Silvia De Nicolò*[†] Enrico Fabrizi[‡] Aldo Gardini[†]

[†]Università di Bologna
[‡]Università Cattolica del S. Cuore

Abstract

The paper targets the estimation of the 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. In particular, 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.

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

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 labelled as *small areas* or *domains*, is particularly useful. When the domain-specific sample sizes

*silvia.denicolo@unibo.it

are too small, the precision of survey estimates is not adequate. Small area estimation (SAE) models 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. The need for SAE techniques emerges since samples available at the upazila level are often very small and, for more than 30% of the areas, no observations are recorded.

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 et al., 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 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 et al., 2013). In this spirit, we integrate auxiliary information taken from remote sensing (RS), as has already been considered for poverty mapping in previous researches (Engstrom et al., 2017; Masaki et al., 2020; Steele et al., 2017). Differently from Schmid et al. (2017) and Steele et al. (2021), we do not make use of mobile operator call detail records (CDR) or other big data sources. However, Steele et al. (2017) note that, when estimating WI-based poverty in Bangladesh, the CDR explanatory variables do not add valuable information with respect to those provided by RS regressors. The set of variables we identify covers different determinants of poverty such as the population structure and density, along with geographical, land use, social and economic features.

The areal nature of auxiliary information drives our choice to a model defined at the area level, relating area-specific survey estimates to auxiliary information available at the same level of aggregation. The alternative unit-level models (see, e.g., Molina et al., 2014) exploit auxiliary variables specified at the unit or household level for the whole population, not available in this case.

We remark that the target measure is a proportion and it is defined on the unit interval. In this respect, the area level literature relies on two main approaches: linear mixed models (Marhuenda et al., 2013), possibly specified on arc-sine transformations (Casas-Cordero Valencia et al., 2016; Schmid et al., 2017), or Beta mixed models. In this paper, we opt for the latter approach as the Beta distribution is intrinsically defined on the $(0, 1)$ interval and allows for asymmetric sampling distributions, common when estimating rates in small samples. The inferential setting we adopt is the Bayesian one; among earlier contributions relevant to our research we mention Fabrizi et al. (2011); Janicki (2020); Liu et al. (2007). Adopting a Hierarchical Bayes (HB) approach has several benefits in the small area estimation context (Rao and Molina, 2015, Chapter 10), especially those of easily managing non-Gaussian distributional assumptions and fully capturing the uncertainty around target parameters.

The framework described so far comprises some methodological issues. Firstly, when estimating proportions from samples with small sizes, the possibility to observe all 0 or 1 values cannot be ruled out. This may impact the way of modelling direct estimators since the Beta model does not handle such values. Ad-hoc solutions for this issue have been proposed, e.g., by Wieczorek and Hawala (2011), Fabrizi et al. (2016) and Fabrizi et al. (2020). Secondly, the DHS data are characterized by strong intra-cluster correlation, as already pointed out by Schmid et al. (2017) for Senegal DHS. This makes direct estimators less efficient and increases the probability of observing proportions equal to either 0 or 1. Thirdly, the potentially large number of covariates to be included in the model poses a problem of variable selection. Eventually, the high level of spatial disaggregation leads to possible areas without observations for which we need to provide reliable predictions.

Our proposal contributes to the literature in different directions. We started from Fabrizi et al. (2016), extending their model to account for the observation of 1 values. In doing this, we keep their assumption that direct estimates equal to either 0 or 1 are due to a censoring process, namely, such limit values are only observed because of reduced area-specific sample sizes. The true population values may be very close but not exactly equal to either 0 or 1. Nonetheless, we face the strong clustering effect characterizing survey data by providing a substantial model generalization that relaxes their independence assumption in modelling 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, 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. With the spirit of providing a method that can be widely applied by final users, we implement a set of flexible prior choices that do not need fine tuning interventions. In this line, the use of the horseshoe prior for regression coefficients is complemented by a type of spike-and-slab prior for the random effects (Fabrizi et al., 2018; Tang et al., 2018). Lastly, we propose a new methodology for out-of-sample predictions that propagates the related uncertainty.

We assess the frequentist properties of the proposed predictor using a design-based simulation in comparison with existing models in the literature, namely the Fay-Herriot with arc-sine transformation and, among those relying on the Beta likelihood, the one by Fabrizi et al. (2016). We find that the proposed predictors 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 et al. (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 standardised. 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: $[0; 1)$, $[1; 5), [5; 10)$, $[10; 15), \dots, [75; 80)$, $[80; +\infty)$, 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 being acknowledged to be a proxy of economic development (Masaki et al., 2020; Zhou et al., 2015). 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 behaviours 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 characterising the productivity of the area.

3 Poverty estimator

In this section, we introduce the direct estimator $\hat{\hat{Y}}_d$ of the head-count 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

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

with D_s being the number of in-sample upazilas. The estimator $\hat{\hat{Y}}_d$, suitable for the estimation of the mean in unplanned domains, 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 to be known which 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.

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.

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 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 et al. (1999) and adjusted by Lynn et al. (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 et al. (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 using:

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

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 a 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 et al. (2016), is fully described in Section 4.1. An alternative approach relying on the classical Fay-Herriot model with the arc-sine transformation of the direct estimates (Schmid et al., 2017) is presented in Section 4.2.

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 $\mathbb{V}(y) = \mu(1-\mu)(\phi+1)^{-1}$. For this reason, when modelling proportions, $\phi+1$ can be interpreted as an 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 area-specific covariates, $\boldsymbol{\beta}$ the vector of regression coefficients, v_d an area-specific random effect and ϕ_d the area-specific dispersion parameter, usually assumed to be known for guaranteeing identifiability such as in this case.

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):

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

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 boundaries 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 modelling π_{0d} and π_{1d} resembles the one of Fabrizi et al. (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 observe 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 et al. (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

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

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

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

Before we turn to prior specification, we note that the parameter ϕ_d is assumed known, in line with many small area estimation applications. 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. Note from (8) that $\mathbb{V}\left[\hat{\hat{Y}}_d | \mu_d, \pi_{0d}, \pi_{1d} \right]$ depends on F_d only through the first addend related to the occurrence $\hat{\hat{Y}}_d \in (0, 1)$.

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: $\mathbf{z} = \{\text{logit}(\hat{\hat{Y}}_d)\}$, $d \in D_s$ 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 et al., 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 et al. (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 behaviour of a spike-and-slab prior. More in detail, we specify:

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

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 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 $\hat{\theta}_d^{HB} \xrightarrow{p} \hat{\bar{Y}}_d$. 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 $\boldsymbol{\beta}|\text{data}$ along with one from $v_d|\text{data}$. As v_d constitutes a random effect from an unobserved area. Having the b-th Monte Carlo replicate from the posterior distribution $\xi|\text{data}$, 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 an alternative model, commonly used in the case of small area estimation of ratios and proportions, and namely the Fay-Herriot model with arc-sine square root transformation. This model is adopted by Raghunathan et al. (2007), Casas-Cordero Valencia et al. (2016) in the context of poverty mapping, and by Schmid et al. (2017) among others. Frequentist prediction is implemented in the `emdi` R package (Kreutzmann et al., 2019). 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, i.e. $S_d^2 \cong 1/(4\tilde{n}_d)$ (Efron and Morris, 1975).

We propose a HB approach for estimating the arc-sine square root 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.

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) and Arc-Sine (AS) models. We also introduce in the comparison the model by Fabrizi et al. (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 FH model with arc-sine transformation (EBLUP-AS) provided by the package `emdi`;

2. The Hierarchical Bayes (HB) estimator under FH model with arc-sine transformation (HB-AS);

3. The HB estimator under the Extended Beta model (HB-EB);

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

4. 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-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 modelling 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. The considered small-area models behave rather similarly. As expected, the direct estimator is unbiased, and 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. On 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 HB-AS and HB-FFT in approximately 7 out of 10 areas (70.3%). The left plot of Figure 1

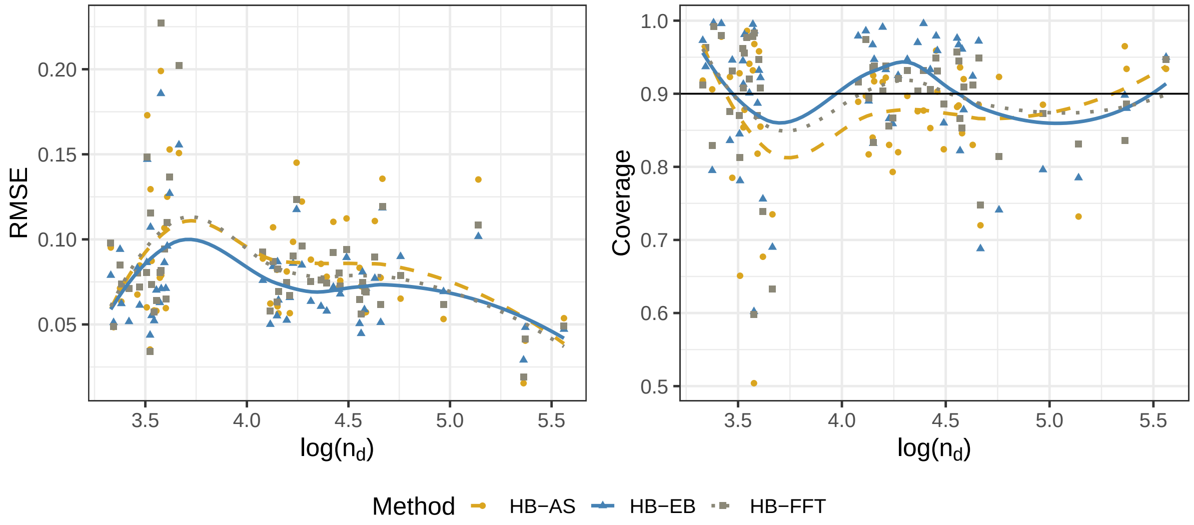

Figure 1: Behaviour of RMSE and frequentist coverage with respect to the area sample size n_d.

shows the behaviour 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 AS-EB and HB-FFT models, which surprisingly behave similarly. 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 1. 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 behaviour 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. A first possible 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.

	EB Model				FFT Model			
	Post. Mean	Post. SD	$Q_{0.025}$	$Q_{0.975}$	Post. Mean	Post. SD	$Q_{0.025}$	$Q_{0.975}$
ζ	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 ζ and λ and LOOIC.

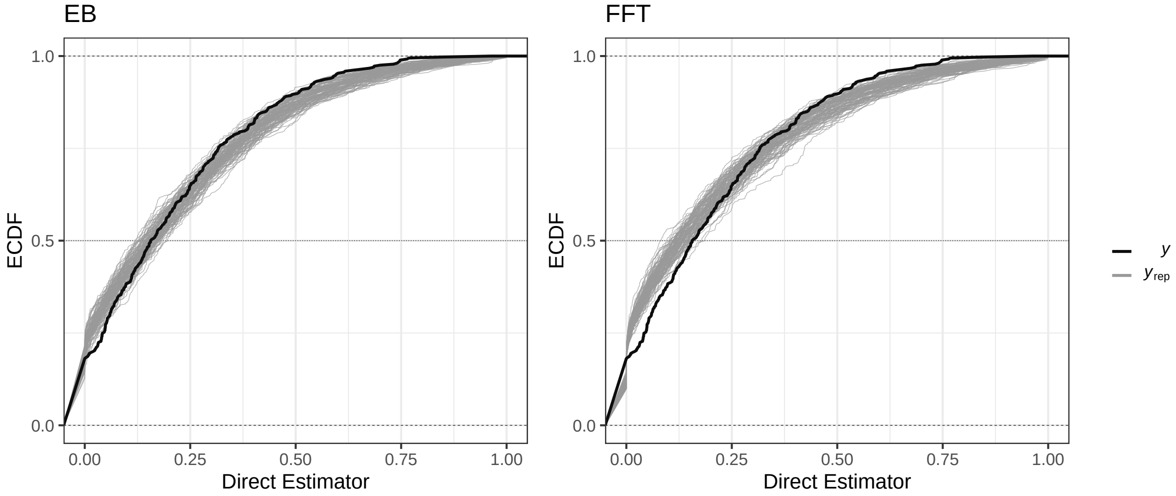

Figure 2: Empirical cumulative distribution functions (ECDF) of the posterior predictive distributions under models EB and FFT.

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), with 66 zero values, while 179 areas are out-of-sample.

We carry on estimation only on EB and FFT models. This is mainly due to the nested assumptions characterising the two models, which allows for the employment of model selection tools able to drive a clear comparison. The arc-sine model is ruled out from the analysis since simulation results do not point out higher performances. Furthermore, the different likelihood assumptions and link functions would compound the model comparison. The Beta-based models offer the additional pro of favouring the interpretability of regression coefficients: they are on the logit scale and, once exponentiated, they can be read in terms of probability odds.

The horseshoe priors described in Section 4.1.1 needs 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. Lastly, the data pseudo-variance resulted to be $\tilde{\sigma}^2 = 1.51$ by applying (10).

As regards model comparison, both the leave-one-out information criteria (LOOIC, Vehtari et al., 2017) and the posterior predictive checks (Gabry et al., 2019) point out

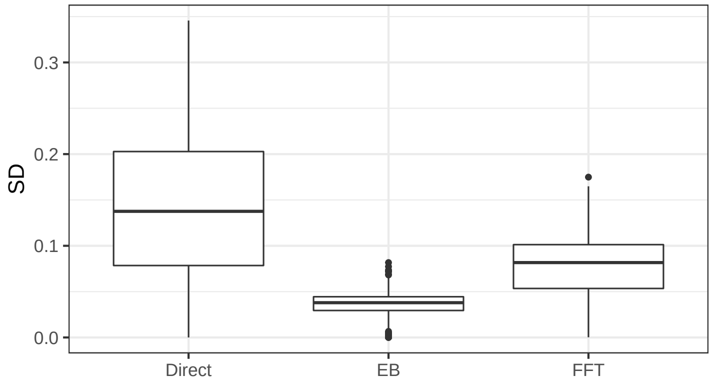

Figure 3: Distributions of the posterior standard deviations of HB estimates under models EB and FFT compared to the standard errors of Direct estimates.

Covariate	Transformation	$\mathbb{E}[\beta_j \| \text{data}]$	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

that the introduction of the correlation parameter λ substantially improves model performances. In detail, Table 3 shows lower values of LOOIC concerning the EB model. In Figure 2, we compare the empirical cumulative distribution function (ECDF) related to the original sample (in black) and the ones related to samples generated under the models (in grey). This posterior predictive check highlights the inability of FFT to model the probabilities of the censored values, leading to a systematic underestimation. Focusing on small area diagnostics, the standard deviations of model-based estimators are remarkably lower than those related to direct estimators (Figure 3). Specifically, the EB model ones are more reliable than the FFT ones.

By considering the posterior summaries in Table 3, we note that the misspecification in modeling 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. Focusing on the correlation parameter λ, we observe that is well identified by the data, having 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.

Table 4 reports the posterior summaries of regression coefficients for the most impor-

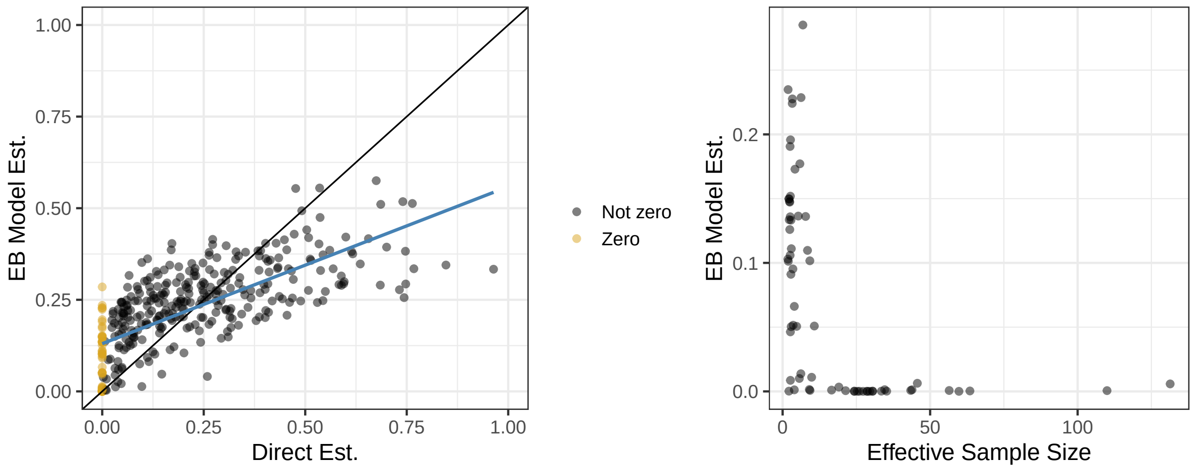

Figure 4: Left plot: comparison between EB model estimates and direct estimates (linear regression line reported). Right plot: focus on areas with 0 direct estimates, comparing model-based estimates to the effective sample size.

tant 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 et al., 2017), remoteness and exclusion from the national labour and goods market represent one of the main drivers of poverty in the community level in Bangladesh.

The amount of shrinkage induced by the model is described in the left panel of Figure 4, 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 4 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 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 (Islam et al., 2017; Khudri et al., 2013) as far as wealth indicators are concerned.

A map of poverty estimates at the upazila level can be found in Figure 5 as regards both direct and model-based methods. If we look at the map, we see how model predictions fill the many gray areas (out-of-sample), especially in peripheral regions. It is important to note that the maps share the same gradient scale, highlighting the shrinking process induced by the model from a spatial viewpoint. Nonetheless, some gaps 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

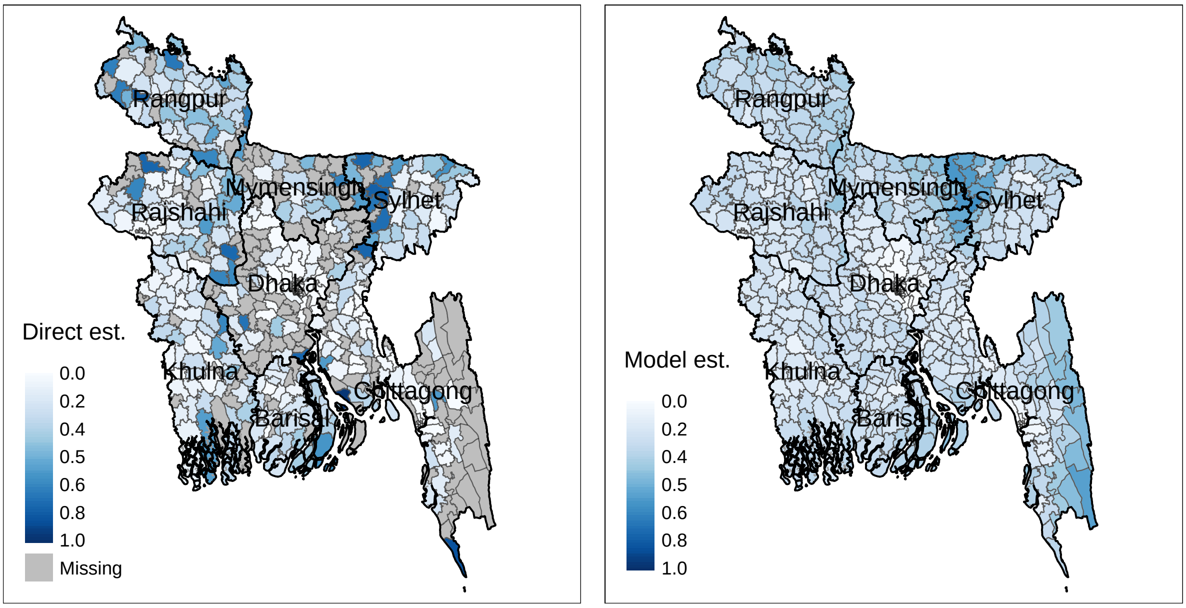

Figure 5: Map with direct estimates and model-based estimates for the Bangladesh upazilas.

have the highest, e.g. the easternmost upazilas near the Indian/Burma border. A cluster of upazilas characterized by high poverty rates shows up in the North-East (Sylhet basin lowlands): they are particularly exposed to climate change effects and floodings (Haque and Jahan, 2015). 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).

Figure 6 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.

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 prior specification. Indeed, the latter aspect has often represented a limit to the widespread use of Bayesian methods among practitioners. Furthermore, our methodology is going to be released in the R `tipsae` package (De Nicolò and Gardini, 2022), 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 et al. (2016) to more effectively handle data

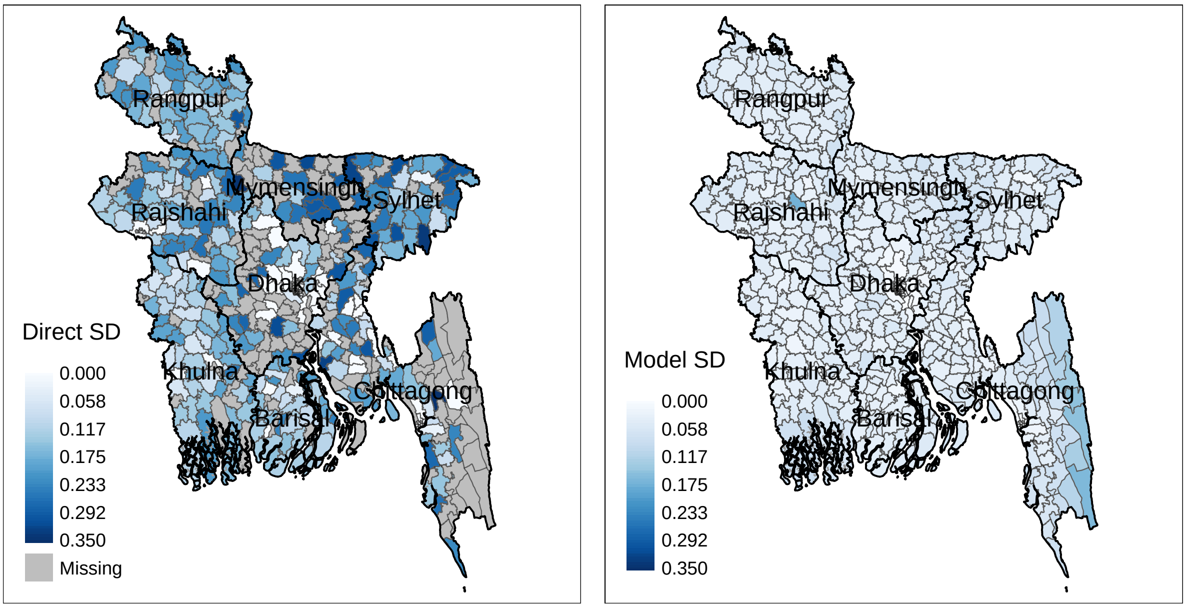

Figure 6: Standard error of direct estimates and posterior standard deviation of model-based estimates.

features as the presence of estimates equal to either 0 or 1 and the strong intra-cluster correlation of observations. The simulation and application results underline the 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).

Our research 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.

References

P. J. Brown and J. E. Griffin. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

C. R. Burgert, B. Zachary, and J. Colston. Incorporating geographic information into demographic and health surveys: a field guide to gps data collection. *Calverton, Maryland, USA: ICF International*, 2013.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

C. Casas-Cordero Valencia, J. Encina, and P. Lahiri. Poverty mapping for the chilean comunas. *Analysis of Poverty Data by Small Area Estimation*, pages 379–404, 2016.

S. Chen and K. Rust. An extension of kish's formula for design effects to two-and three-stage designs with stratification. *Journal of Survey Statistics and Methodology*, 5(2): 111–130, 2017.

D. J. Corsi, M. Neuman, J. E. Finlay, and S. Subramanian. Demographic and health surveys: a profile. *International journal of epidemiology*, 41(6):1602–1613, 2012.

G. S. Datta, P. Hall, and A. Mandal. Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106(493):362–374, 2011.

S. De Nicolò and A. Gardini. *tipsae: Tools for handling Indices and Proportions in Small Area Estimation*, 2022. R package version 0.0.4.

G. Duranton and A. J. Venables. Place-based policies: principles and developing country applications. In *Handbook of regional science*, pages 1009–1030. Springer, 2021.

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

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

E. Fabrizi, M. R. Ferrante, S. Pacei, and C. Trivisano. Hierarchical bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Computational Statistics & Data Analysis*, 55(4):1736–1747, 2011.

E. Fabrizi, M. Ferrante, and C. Trivisano. 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*, pages 299–314, 2016.

E. Fabrizi, M. R. Ferrante, and C. Trivisano. Bayesian small area estimation for skewed business survey variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4):861–879, 2018.

E. Fabrizi, M. R. Ferrante, and C. Trivisano. 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(3):1273–1291, 2020.

S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

W. A. Fuller. *Sampling Statistics*. John Wiley & Sons, 2011.

S. Gabler, S. Häder, and P. Lahiri. A model based justification of kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25:105–106, 1999.

J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, A. Gelman, et al. Visualization in bayesian workflow. *Journal of the Royal Statistical Society Series A*, 182(2):389–402, 2019.

J. Hájek. Discussion of 'an essay on the logical foundations of survey sampling, part i', by d. basu. *Foundations of Statistical Inference*, page 326, 1971.

A. Haque and S. Jahan. Impact of flood disasters in bangladesh: A multi-sector regional analysis. *International Journal of Disaster Risk Reduction*, 13:266–275, 2015.

S. I. Hay and R. W. Snow. The malaria atlas project: developing global maps of malaria risk. *PLoS medicine*, 3(12):e473, 2006.

A. Iimi, F. Ahmed, E. C. Anderson, A. S. Diehl, L. Maiyo, T. Peralta-Quirós, and K. Rao. New rural access index: main determinants and correlation to poverty. *World Bank Policy Research Working Paper*, (7876), 2016.

D. Islam, J. Sayeed, and N. Hossain. On determinants of poverty and inequality in bangladesh. *Journal of Poverty*, 21(4):352–371, 2017.

R. Janicki. 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(9):2264–2284, 2020.

G. Kalton. Ultimate cluster sampling. *Journal of the Royal Statistical Society: Series A (General)*, 142(2):210–222, 1979.

M. M. Khudri, F. Chowdhury, et al. Evaluation of socio-economic status of households and identifying key determinants of poverty in bangladesh. *European Journal of Social Sciences*, 37(3):377–387, 2013.

L. Kish. Weighting in deft2. *The Survey Statistician*, 17(1):26–30, 1987.

J. Klotz. Statistical inference in bernoulli trials with dependence. *The Annals of statistics*, pages 373–379, 1973.

A.-K. Kreutzmann, S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis. The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91(7):1–33, 2019. doi: 10.18637/jss.v091.i07.

B. Liu, P. Lahiri, and G. Kalton. Hierarchical bayes modeling of survey-weighted small area proportions. In *Proceedings of the American Statistical Association, Survey Research Section*, pages 3181–3186, 2007.

P. Lynn, S. Häder, and S. Gabler. Design effects for multiple design samples. *Survey Methodology*, 32(1):115–120, 2006.

Y. Marhuenda, I. Molina, and D. Morales. Small area estimation with spatio-temporal fay-herriot models. *Computational Statistics & Data Analysis*, 58:308–325, 2013.

T. Masaki, D. Newhouse, A. R. Silwal, A. Bedada, and R. Engstrom. Small area estimation of non-monetary poverty with geospatial data. 2020.

I. Molina, B. Nandram, and J. Rao. Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. *The Annals of Applied Statistics*, 8(2):852–885, 2014.

NIPORT and Mitra and Associates and ICF International. 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, 2016.

M. S. O'Donnell and D. A. Ignizio. Bioclimatic predictors for supporting ecological applications in the conterminous united states. *US geological survey data series*, 691 (10):4–9, 2012.

C. Pezzulo, G. M. Hornby, A. Sorichetta, A. E. Gaughan, C. Linard, T. J. Bird, D. Kerr, C. T. Lloyd, and A. J. Tatem. Sub-national mapping of population pyramids and dependency ratios in africa and asia. *Scientific Data*, 4(1):1–15, 2017.

J. Piironen and A. Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018 – 5051, 2017. doi: 10.1214/17-EJS1337SI.

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

T. E. Raghunathan, D. Xie, N. Schenker, V. L. Parsons, W. W. Davis, K. W. Dodd, and E. J. Feuer. Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102(478):474–486, 2007.

M. Rahman et al. Role of agriculture in bangladesh economy: uncovering the problems and challenges. *International Journal of Business and Management Invention*, 6(7), 2017.

J. N. Rao and I. Molina. *Small area estimation.* John Wiley & Sons, 2015.

M. S. Ridout, C. G. Demetrio, and D. Firth. Estimating intraclass correlation for binary data. *Biometrics*, 55(1):137–148, 1999.

T. Schmid, F. Bruckschen, N. Salvati, and T. Zbiranski. 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(4):1163–1190, 2017.

J. E. Steele, P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. De Montjoye, A. M. Iqbal, et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127): 20160690, 2017.

J. E. Steele, C. Pezzulo, M. Albert, C. J. Brooks, E. zu Erbach-Schoenberg, S. B. O'Connor, P. R. Sundsøy, K. Engø-Monsen, K. Nilsen, B. Graupe, et al. Mobility and phone call behavior explain patterns in poverty at high-resolution across multiple settings. *Humanities and Social Sciences Communications*, 8(1):1–12, 2021.

S. Sugasawa and T. Kubokawa. Transforming response values in small area prediction. *Computational Statistics & Data Analysis*, 114:47–60, 2017.

X. Tang, M. Ghosh, N. S. Ha, and J. Sedransk. Modeling random effects using global–local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 113(524):1476–1489, 2018.

A. J. Tatem. Worldpop, open data for spatial demography. *Scientific Data*, 4(1):1–4, 2017.

A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017.

D. I. Warton and F. K. Hui. The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 92(1):3–10, 2011.

J. Wieczorek and S. Hawala. 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*, 2011.

World Bank. *Poverty Assessment for Bangladesh: Creating Opportunities and Bridging the East-West Divide*. World Bank, 2008.

Y. Zhou, T. Ma, C. Zhou, and T. Xu. Nighttime light derived assessment of regional inequality of socioeconomic development in china. *Remote Sensing*, 7(2):1242–1262, 2015.