Journal of the Royal Statistical Society Series C: Applied Statistics, 2025, **74**, 1255–1278
https://doi.org/10.1093/jrsssc/qlaf022
Advance access publication 26 March 2025
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

A Mixture-of-Experts model to deal with the rural/urban dichotomy in small area estimation

Aldo Gardini[1], Silvia De Nicolò[1] and Enrico Fabrizi[2]

[1]Department of Statistical Sciences, University of Bologna, Bologna 40126, Italy
[2]DISES, Catholic University of the Sacred Heart, Piacenza 29122, Italy

Address for correspondence: Silvia De Nicolò, Department of Statistical Sciences, University of Bologna, via Belle Arti 41, Bologna (BO) 40126, Italy. Email: silvia.denicolo@unibo.it

Abstract

Small area estimation is a pivotal statistical methodology for generating reliable estimates of socio-economic indicators across small sub-populations. This research focuses on estimating the per capita wealth index in Bangladeshi upazilas, by integrating data from the Demographic and Health Survey with remote sensing covariates. The Fay–Herriot model, a prominent area-level model in small area literature, shows limitations when applied to our data. Notably, an exploratory analysis unveils two distinct regimes in the data-generating process, characterized by the rural/urban dichotomy. To tackle this issue, we extend the Fay–Herriot model through a Mixture-of-Experts, which classifies areas into groups within the estimation process. This enhances flexibility while preserving desirable properties, such as design consistency and interpretability of predictors. In addition, this family of models defines the mixing probabilities through a logistic regression, which proved to be particularly efficient in our applied setting, as demonstrated by a simulation study.

Keywords: Bayesian inference, DHS survey, Fay–Herriot model, wellbeing indicators

1 Introduction

The need for estimates of socio-economic indicators for specific sub-populations is growing, with the aim of monitoring the geographical or socio-demographic distribution of poverty, inequality, and wealth. In this context, the main data sources are population surveys that collect information on features such as income, consumption, and well-being at the individual/household level. When the target is a large set of sub-populations, survey data alone typically falls short of providing reliable estimates of indicators of interest due to restricted sample sizes. To produce reliable estimates for such sub-populations, it is possible to rely on the so-called small area estimation (SAE) techniques that integrate survey data with available auxiliary information. Such alternative data sources may include Censuses, administrative registers, satellite images, and phone traffic data; their integration with survey data can be performed either at the unit/household or area level (see Rao & Molina, 2015, Chapters 6 and 7). In this context, the terms *area* or *domain* denote the target sub-population, although it is not necessarily a geographical region.

The Fay–Herriot (FH, Fay & Herriot, 1979) model is the simplest and most popular area-level model. One distinguishing feature of this model is the linear regression which links the area-level target parameter to the auxiliary information. This entails the assumption of a single intercept and regression slopes for the entire ensemble of areas considered. Such an assumption can be very restrictive when studying problems characterized by a relationship between response and covariates

Received: February 15, 2024. **Revised:** December 20, 2024. **Accepted:** March 4, 2025
© The Royal Statistical Society 2025.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

that vary across areas. For example, this could happen when targeting heterogeneous populations or when the target indicator is governed by distinct generative processes, as in the case of high socio-economic or territorial stratification (Diz-Rosales et al., 2024).

In this article, the need for an extension of the FH model is motivated by a real-data application. Specifically, we consider the estimation of the per capita wealth index (WI) in Bangladeshi upazilas, i.e. small administrative subdivisions similar to boroughs. Direct survey estimates are obtained from the Demographic and Health Survey (DHS), while auxiliary information comes mostly from remote sensing (RS) satellite images. Notably, the WI, as defined in the DHS framework, represents a popular choice in poverty mapping exercises (see, for instance, Chi et al., 2022; Lee & Braithwaite, 2022; Steele et al., 2017; Yeh et al., 2020). Indeed, the DHS provides the WI at the household level to monitor socioeconomic status in over 81 countries worldwide (Rutstein & Johnson, 2004); thus, the relevance of our analysis goes beyond Bangladesh. A rural/urban divide in the WI distribution is typically observed for developing countries, induced by markedly different economies (Rigg, 1998). From a statistical perspective, this can translate into the existence of separate data-generating processes to be considered. This step could be non-trivial in area-level models, as the rural/urban classification of territories could be unknown or may not perfectly overlap with administrative boundaries, as in our case. Therefore, the areas cannot be classified a priori.

We address this challenge by proposing a Bayesian Mixture-of-Experts (MoE) model (Chapter 12 of Frühwirth-Schnatter et al., 2019; Nguyen & Chamroukhi, 2018) that extends the FH model by allowing for the existence of different groups of areas, each characterized by distinct predicting equations. An important feature of MoE is modelling the area-specific probability of belonging to a given group (also known as *gating* probabilities) through logistic regression. Each model component (i.e. the *expert*) is a distinct FH model, allowing for different predictors, coefficients, and area-specific random effects.

The proposed framework is related to previous proposals in the SAE literature. A first body of literature deals with the different relationships between response and covariates through random slope models (Diz-Rosales et al., 2024; Hobza & Morales, 2013; Prasad & Rao, 1990), noting that only the latter contribution refers to area-level models. A second strand relies on clustering procedures: Maiti et al. (2014) combine the FH model with a model-based clustering for areas; whereas, Torkashvand et al. (2017) propose a clustering of areas and a subsequent estimation of separate FH models on clusters. Lastly, we also mention area-level models based on mixtures. The authors of Chakraborty et al. (2016) set a mixture of two Gaussian distributions for random effects; Polettini (2017) extends this framework by means of a Dirichlet process prior on the number of components; Gershunskaya and Savitsky (2020) specify a mixture allowing for group-specific intercepts and variance components, but with common slopes. However, such proposals consider cluster-specific probabilities as constant across areas, as their main focus is on handling departures from normality.

The main novelty of our proposal to small area modelling is related to the way we model membership probabilities, which are allowed to vary across areas. Our focus is to exploit the classification of a specific area to link it probabilistically to the most appropriate predictor, accounting for its peculiarities. Our modelling framework leads to further benefits. Firstly, the MoE approach allows for flexibility in the modelling process while retaining important properties associated with the FH model, such as design consistency and easy interpretation of the predictors. Secondly, our proposal incorporates both clustering and small area estimation steps into a single procedure. As such, it inherently accounts for the uncertainty involved in the clustering step in the resulting estimator. Moreover, the MoE provides a distinct advantage when prediction has to be performed for out-of-sample (OOS) areas, as allowing for stochastic group probabilities can be useful in this step. Such an aspect is particularly relevant in our application as 33% of upazilas are OOS and require prediction. These strengths are confirmed by a simulation study and by our real-data application, in which our model outperforms the FH model and some alternative specifications.

The remainder of this article is structured as follows. Section 2 is devoted to introducing DHS data, basic notations, and relevant survey quantities. Section 3 illustrates the RS covariates, together with an exploratory analysis as a motivating exercise. The FH model and its extension through MoE are described in Section 4. A model-based simulation study is presented in Section 5, and results from a real-data application on WI estimation in Bangladesh are commented in Section 6. Finally, concluding remarks are outlined in Section 7.

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

2 Survey and notation

In this section, we introduce the DHS data that motivate our study and some interesting survey-based quantities employed in the modelling proposal of Section 4. The DHS is introduced in Section 2.1, while basic notation and relevant survey-based quantities are described in Section 2.2.

2.1 The demographic and health survey

DHS are nationally representative household surveys conducted to monitor population health and living conditions in low- and middle-income countries (Corsi et al., 2012). Their key advantages include high response rates, national coverage, and high-quality interviewer training. In the case of Bangladesh DHS, the survey targets the entire population residing in non-institutional dwelling units and is designed to produce representative results for the country as a whole, for the urban and rural areas separately and for each of the seven administrative divisions. Such divisions can be further disaggregated into 64 districts called zilas and 544 sub-districts called upazilas, comparable with counties or boroughs, which are our target areas. The latter ones are unplanned domains with unreliable direct estimates, requiring the use of SAE techniques.

The survey has a two-stage stratified sampling scheme. In the first stage, 600 primary sampling units (PSUs), constituted by census tracts, are selected with unequal probability depending on the PSU size within 20 strata. The strata result from the combination of the seven administrative divisions and a territorial classification that divides the areas into rural, city corporations, and other urban areas. Among the sampled 600 PSUs, 35% are classified as urban or city corporations, and the remaining 65% are classified as rural. In the second stage, 30 households are drawn from every sampled PSU with equal probability. Because of the unequal probabilities of selection and non-response, each household is weighted and survey weights are released along with survey data (NIPORT and Mitra and Associates and ICF International, 2016).

To clarify the terminology, we distinguish between survey design elements, such as strata, PSUs, and households as secondary sampling units, and non-survey elements, such as the administrative units (upazilas). Notably, upazilas do not nest within survey strata and may enclose multiple PSUs. In the following, we refer to upazilas as areas or domains; whereas, when referring to specific groups of upazilas, we denote them as groups or clusters. Our DHS data refer to the 2014 survey wave, recording 18,000 residential households and 81,624 individuals overall. We observe 365 upazilas that include at least one sampled PSU, while 179 upazilas are OOS. By considering in-sample (IS) upazilas, the individual sample sizes span from 16 to 1,884, with a median of 160, denoting a large heterogeneity in sizes and a large number of small domains.

The WI is provided as a survey variable in DHS data, being constructed from a set of questions on household durable assets and housing characteristics such as floor type and ceiling material, toilet or latrine access and phone ownership. Given such a set of variables, the index is constructed by extracting a common factor explaining the largest percentage of the total variance via principal component analysis (PCA) and then adjusting for differences in urban and rural strata. For a discussion of the economic and welfare characterization of the WI with respect to income or consumption, see Poirier et al. (2020). In line with the literature, we target our analysis at the individual level: thus, all individuals belonging to the same household are assumed to share the same WI score.

2.2 Notation and survey-based quantities

The application that motivates this research is the estimation of poverty in Bangladeshi upazila based on survey data. To set some notation, $\mathcal{U}$ indicates a finite population composed of N individuals partitioned into D sub-populations $\mathcal{U}_d$ that determine the target areas. Each sub-population $\mathcal{U}_d$ includes N_d units, such that $N = \sum_{d=1}^{D} N_d$. A sample of overall size n is drawn from $\mathcal{U}$ through a complex survey scheme. As observations are not available for all areas defined by the partition, we order the areas such that the D_{IS} ones, for which at least one unit is observed, appear first followed by the $D_{\text{OOS}} = D - D_{\text{IS}}$ OOS areas. The sample sizes are denoted by n_d for each IS area $d = 1, \ldots, D_{\text{IS}}$ so that $n = \sum_{d=1}^{D_{\text{IS}}} n_d$.

The quantity we target is the WI, as described in Section 2.1; specifically, we are interested in per capita WI in each domain, defined as $\bar{Y}_d = N_d^{-1} \sum_{j=1}^{N_d} y_{dj}$; where y_{dj} denotes the observed WI of

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

individual j belonging to area d. The Hájek-type estimator (Hájek, 1971) is proposed as the direct estimator of $\bar{Y}_d$, as the domains we target are typically unplanned domains, and it is defined as

$$\hat{\bar{Y}}_d = \frac{\sum_{j=1}^{n_d} w_{dj} y_{dj}}{\sum_{j=1}^{n_{dj}} w_{dj}}, \quad d = 1, \ldots, D_{\text{IS}}, \tag{1}$$

where w_{dj} is the sampling weight associated with each individual. Such an estimator is approximately unbiased for the unknown population mean $\bar{Y}_d$.

When estimating the sampling variance of $\hat{\bar{Y}}_d$ under complex two-stage sampling designs, a standard approach is the Ultimate Cluster technique (Kalton, 1979). However, this approach cannot be implemented at the upazila level as a very low number of PSUs is available for each upazila: 86% of IS areas record observations pertaining only to one or two PSUs. Therefore, we opt for a different approach by relying on the estimation of design effects at a higher level of aggregation (zila level). The features to be accounted for in the computation of design effects are the unequal sampling weights and clustering within PSUs. Thus, we adopt the formula by Kish (1992) within each zila, by blending the weights and PSU components. As adapted from Gabler et al. (1999), the formula is defined as

$$\text{DEff}_s = \left[1 + \text{cv}^2(\mathbf{w}_s)\right]\left[1 + \left(\frac{\sum_{i=1}^{c_s}\left(\sum_{j=1}^{n_i} w_{ij}\right)^2}{\sum_{i=1}^{c_s}\sum_{j=1}^{n_i} w_{ij}^2} - 1\right)\rho_s\right], \tag{2}$$

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-PSU correlation coefficient, c_s being the number of PSUs in s, n_i the units within PSUs i, and w_{ij} the individual weight. The sampling variance under complex design $\mathbb{V}_{cs}[\hat{\bar{Y}}_d]$ is retrieved by applying the definition of DEff as $\mathbb{V}_{cs}[\hat{\bar{Y}}_d] = \text{DEff}_d \times \mathbb{V}_{srs}[\hat{\bar{Y}}_d]$, with $\mathbb{V}_{srs}[\hat{\bar{Y}}_d]$ denoting the variance under simple random sampling and computed as the population variance of the WI within each upazila over its sample size, in line with Lumley (2011). The obtained design effect estimates highly correlate with the corresponding estimates via Ultimate Cluster, having a Pearson correlation of 0.73. Some contributions exploiting the design effect to estimate the uncertainty of direct estimates are, among others, Liu et al. (2014), Casas-Cordero Valencia et al. (2016), and You and Hidiroglou (2023).

3 Remote sensing data and the rural–urban divide

As auxiliary information to be integrated into the model, we consider RS variables. Rasters related to population density, age and sex structures, nighttime light radiance, built-settlement extent areas, and distance to the nearest cultivated areas have been downloaded from the Worldpop database;[1] while a raster including the time to nearest healthcare site comes from the Malaria Atlas Project database.[2] Relevant quantities have been computed at the upazila level by cutting rasters through shapefiles.

To distinguish between rural, urban, or metropolitan areas, we need to summarize population distribution and its spatial heterogeneity. In this spirit, we consider the average population density and its Geary index. Moreover, we incorporate the demographic composition by age and sex classes, including the male/female and child/woman ratios. The nighttime light radiance and distances to the nearest healthcare site are considered proxies of area urbanization and economic development. As the agricultural sector is a driving force for the Bangladeshi economy, variables related to territorial characterization such as distance to cultivated areas and proportion of built areas are included as further proxies of productivity. A list of the main variables, their original rasters and the aggregation type starting from grid square values is presented in Table 1.

[1] https://hub.worldpop.org/
[2] https://data.malariaatlas.org/

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

Table 1. Remote sensing variables with aggregation type over grid square values and original rasters

Name	Aggregation type	Raster
Pop. Density	Mean of grid square values	Population density
Geary index of Pop. Density	Geary index of grid square values	Population density
Child/Woman	Mean of child over woman of child bearing age (15–49 yo) counts per grid square	Age and sex structures
Male/Female	Mean of male over female counts in each age group per grid square	Age and sex structures
VIIRS	Mean of grid square values	Nighttime light radiance
Prop. Builded Areas	Mean of grid square values	Builded-settlement extent areas
Dist. Cultivated Area	Mean of grid square values	Distance to the nearest cultivated areas
Time to Health Care	Mean of grid square values	Time to nearest healthcare site

3.1 An exploration of the rural–urban divide

When dealing with wealth distribution, a distinction between the rural and urban spheres is usually made (Rigg, 1998). This separation is attributable to intrinsic differences in household economies, with a predominance of primary and secondary sectors in rural Bangladeshi areas and of secondary and tertiary sectors in urban ones (Garrett & Chowdhury, 2004). This translates into a diversification of household assets on which the WI is based, even if harmonized as a single variable. Therefore, it may be reasonable to assume that wealth distribution results from two separate interrelated processes for rural and urban areas. Indeed, rural–urban disparities are also reflected by the RS variables used as covariates, especially those related to the degree of urbanization. As a consequence, the rural–urban divide could imply a different relationship between the WI and RS covariates.

In this framework, an additional issue is that the administrative units under study, i.e. the upazilas, are not characterized by a perfect overlap between rural and urban compositions. Indeed, they represent a combination of both, involving various patterns: mainly rural, metropolitan, or urban centres surrounded by rural zones. The RS rasters can help identify such patterns, as shown in Figure 1. Population density and built-settlement areas are displayed for three different upazilas: Chandgaon, located in the metropolitan area of Chittagong; Belai Chhari, pertaining to the mountain area of the same division; and Bogra Sadar, where the medium-sized town of Bogra is located.

An exploratory clustering analysis relying on RS variables is performed to evaluate their effectiveness in distinguishing between mostly rural and urban upazilas. Such a classification could help us in investigating the existence of separate processes that determine the WI in rural and urban zones. The variables included in this step are those directly related to the degree of urbanization: the mean and Geary index of population density, proportion of built-settlement areas, and nighttime light radiance. We perform a standard hierarchical clustering procedure with complete linkage, even if the results are robust to other linkage criteria (e.g. single and Ward linkages).

Results are displayed by the dendrogram in Figure 2. The optimal number of clusters is selected by maximizing the average silhouette width, being equal to 2. Notably, the three upazilas in Figure 1 are included in the dendrogram of Figure 2 to facilitate interpretation. The clustering procedure identifies two groups of areas, characterized in the plot by different colours: cluster 1, larger in size, comprising peri-urban (such as Bogra Sadar) and predominantly rural areas (such as Belai Chhari); and cluster 2, including metropolitan ones such as Chandgaon. An additional affirmation of result interpretability is evident when cutting the tree into three clusters. In this case, cluster 1 is split into a smaller subset comprising remote locations such as Belai Chhari, and a larger subset encompassing peri-urban areas such as Bogra Sadar. Hence, urbanization patterns seem efficiently detected by our approach.

An interesting cue comes from the left-hand side plot of Figure 3, depicting the distribution of the per capita WI direct estimate versus the two clusters. The level of WI estimates in metropolitan

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

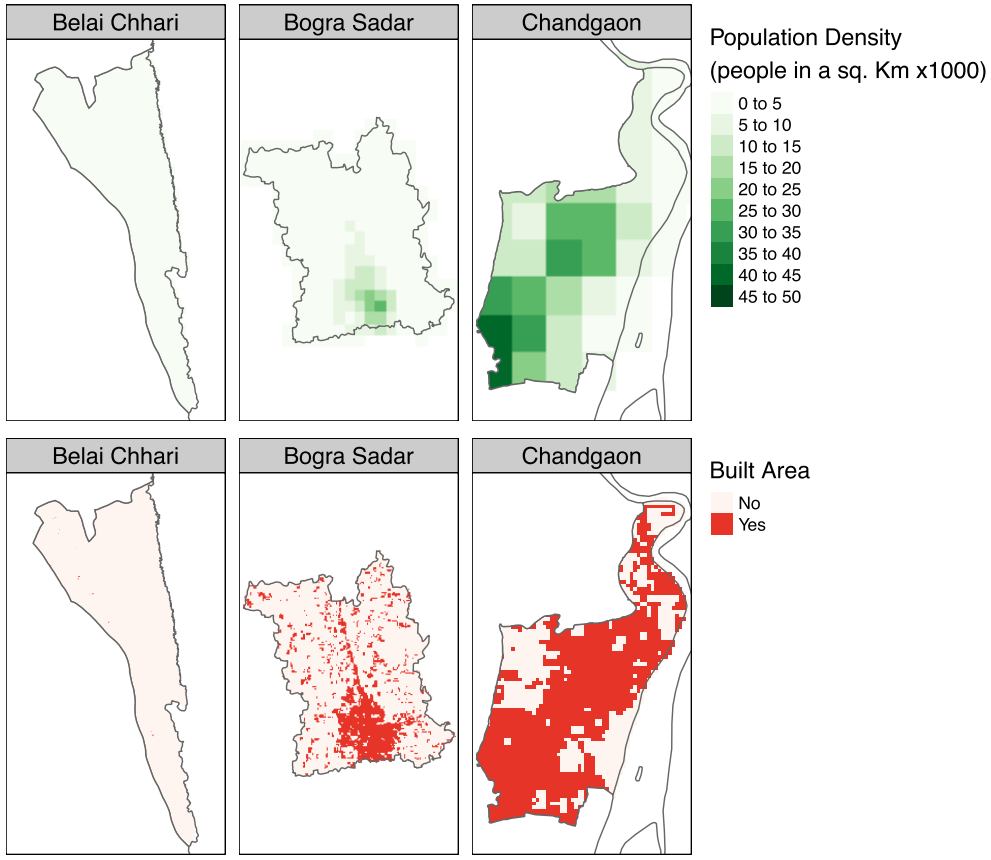


Figure 1. Population density and built-settlement areas displayed for three upazilas showing different patterns.

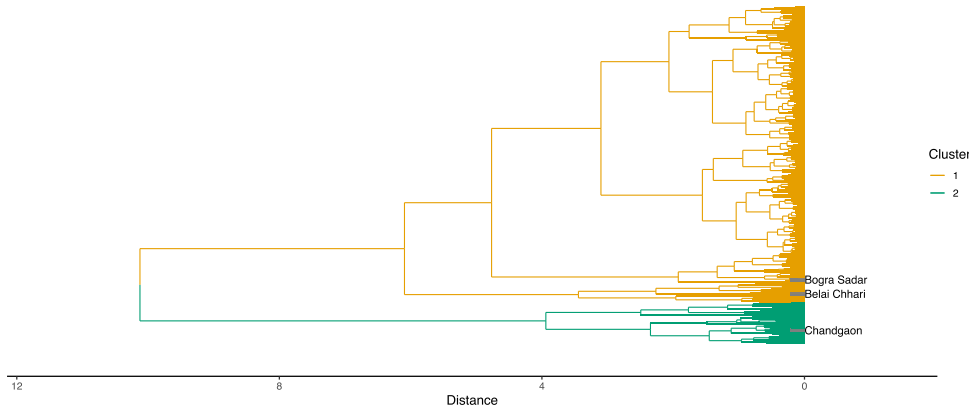


Figure 2. Dendrogram of the Hierarchical Clustering of upazilas based on urbanization variables.

areas appears unquestionably greater than those in peri-urban and predominantly rural areas. This is in line with Bangladeshi poverty trends, with rural areas characterized by higher poverty incidence (Ferdousi & Dehai, 2014) and confirms the fact that wealth distribution may differ between predominantly rural/urban areas.

Downloaded from academic.oup.com/jrssse/article/74/5/1255/8096259 by guest on 28 September 2026

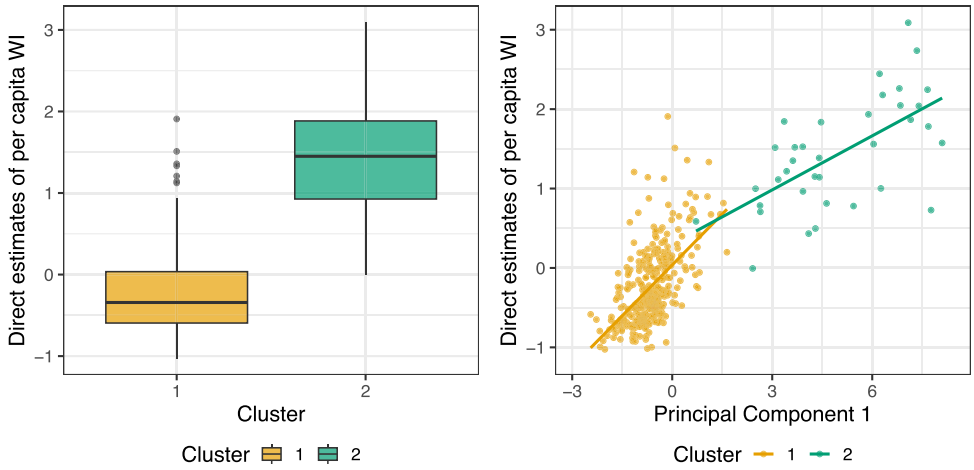


Figure 3. Boxplot of per capita WI direct estimates stratified with respect to two clusters (on the left-hand side), and scatterplot of the first principal component versus per capita WI direct estimates, including linear regression lines (on the right-hand side).

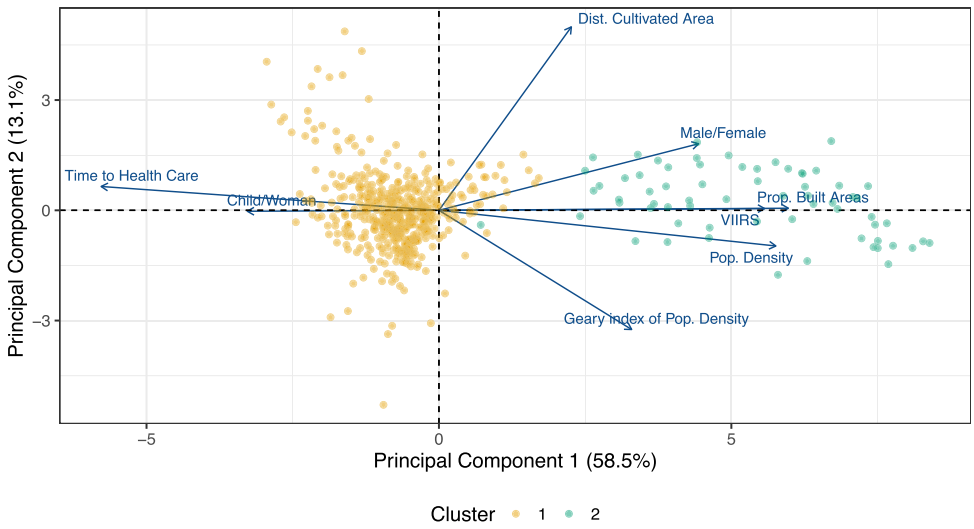


Figure 4. PCA biplot of scores and loadings related to the first two principal components.

The distinct relationships between WI and RS covariates in rural–urban contexts are assessed via PCA. Indeed, the RS variables related to urbanization level appear to be highly correlated, with six pairs of them showing Pearson correlation estimates between 0.80 and 1 in absolute values. Thus, dimensionality reduction seems an easy-to-implement strategy to identify separate trends. The area scores related to the first two principal components, which account for 71.6% of the total variance, are shown in the biplot of Figure 4, together with the loadings. The first principal component (58.5% of total variance) is a clear proxy for the urbanization level: urban areas are characterized by high scores, determined by high values of VIIRS, the proportion of built areas, population density and low time to reach health care sites. The second principal component (13.1% of the total variance) captures another aspect of the territorial structure; it takes high positive values for areas either far from cultivated land or

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

homogeneous in terms of population density. Consequently, high scores are recorded for mountain and uninhabited areas, regarding the rural cluster; and city centre sub-districts, regarding the urban cluster. Conversely, low values are observed for upazilas dotted with villages and surrounded by cultivated fields.

Notably, the first principal component can discriminate the areas belonging to the two clusters. A scatterplot comparing the first component scores versus the direct estimates of per capita WI is shown in the right-hand side plot of Figure 3, with fitted linear regression lines for each cluster. Such lines confirm the different relationships between urbanization degree and per capita WI in upazilas, both in terms of intercepts and slopes. In fact, metropolitan upazilas show a higher baseline wealth level and a lower rate of change in WI with respect to a proxy for urbanization.

In conclusion, the rural–urban divide seems to affect per capita WI and its relationship with RS variables, and such exploratory analysis provides several hints to set up an ad hoc modelling framework. Firstly, we remark that the hierarchical clustering procedure constitutes an attempt to classify areas based on their level of urbanization. However, since several upazilas cannot be assigned to a given cluster straightforwardly, it may be useful to adopt a fuzzy approach by directly modelling the probabilities of cluster membership. Next, the uncertainty deriving from this classification process needs to be considered when separately modelling the WI with respect to the RS variables in distinct clusters. Hence, the next section is devoted to proposing a new small area model that addresses these issues.

4 Fay-Herriot model and its Mixture-of-Experts generalization

In this section, we outline our modelling framework. Specifically, we begin by detailing the standard FH model in Section 4.1, followed by the formalization of our methodological proposal, which extends the FH model, in Section 4.2. To clarify the notation, given our collection of direct estimates $\hat{\bar{Y}}_d$, $d = 1, \ldots, D_{\text{IS}}$, we assume $\mathbb{E}[\hat{\bar{Y}}_d \mid \theta_d] = \theta_d$, where θ_d is the super-population parameter, i.e. the mean of the process generating the finite population, whose mean is $\bar{Y}_d$.

4.1 The Fay-Herriot model

The FH model can be expressed by means of two equations, usually labelled as sampling and linking levels. At the sampling level, a Gaussian distribution is assumed to model the uncertainty associated with direct estimates:

$$\hat{\bar{Y}}_d \mid \theta_d, S_d^2 \overset{ind}{\sim} \mathcal{N}(\theta_d, S_d^2), \quad d = 1, \ldots, D_{\text{IS}}, \tag{3}$$

where the standard error S_d is typically assumed to be known. The unobserved population parameter θ_d is modelled at the linking level under the normality assumption:

$$\theta_d \mid \beta_0, \boldsymbol{\beta}, \sigma_u \overset{ind}{\sim} \mathcal{N}(\beta_0 + \mathbf{x}_d^T \boldsymbol{\beta}, \sigma_u^2), \quad \forall d; \tag{4}$$

where β_0 is the intercept, $\mathbf{x}_d \in \mathbb{R}^P$ contains the values of P auxiliary variables registered for area d, $\boldsymbol{\beta} \in \mathbb{R}^P$ the vector of associated regression coefficients, and σ_u is a scale parameter. Conditioned on the area-specific random effect $u_d \mid \sigma_u \overset{ind}{\sim} \mathcal{N}(0, \sigma_u^2)$, $\forall d$, the target parameter θ_d can also be expressed as follows:

$$\theta_d \mid \beta_0, \boldsymbol{\beta}, u_d = \beta_0 + \mathbf{x}_d^T \boldsymbol{\beta} + u_d, \quad \forall d. \tag{5}$$

The estimation of the FH model can be performed via different approaches, including the Bayesian approach we consider (Rao & Molina, 2015, Chapter 8). To implement a Bayesian inferential procedure, prior distributions must be specified for both model parameters and hyperparameters. A fully proper prior setting is adopted. We specify a standard half-normal distribution for the scale parameter σ_u. Following the guidance of Goodrich

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

et al. (2023), we set the prior for the intercept as $\beta_0 \sim \mathcal{N}(m_y, 2.5^2 s_y^2)$, where m_y and s_y^2 represent the mean and variance of the direct estimates, respectively. Finally, for the regression coefficients, we impose $\beta_p \sim \mathcal{N}(0, 2.5^2), \forall p$, provided that the covariates are standardized.

Note that, in this context, the posterior conditional expectation of θ_d is defined as

$$\mathbb{E}\left[\theta_d \mid \beta_0, \boldsymbol{\beta}, \sigma_u, \hat{\bar{Y}}_d\right] = \hat{\bar{Y}}_d - \eta_d(\hat{\bar{Y}}_d - \beta_0 - \mathbf{x}_d^T\boldsymbol{\beta}) = (1-\eta_d)\hat{\bar{Y}}_d + \eta_d(\beta_0 + \mathbf{x}_d^T\boldsymbol{\beta}), \tag{6}$$

where $\eta_d = S_d^2/(\sigma_u^2 + S_d^2)$. Equation (6) yields what is known as the shrinkage process, which involves combining direct estimates with synthetic estimates to generate more stable and reliable estimates for small areas. This combination is achieved through a weighted approach, where the weight η_d, referred to as the shrinkage factor, is determined by the accuracy of both the direct and synthetic estimates.

4.2 Extending the Fay-Herriot model: Mixture-of-Experts

MoE are suitable models for analysing data whose generating process is decomposable into a set of sub-processes defined on the covariate space (Peng et al., 1996). This family of models has been introduced in the computer science literature, defining a two-level model: the response is modelled assuming a fixed number of components defined as generalized linear models (*expert networks*), and the prior probability of belonging to each group (*gating networks*) is modelled through logistic regression.

In this section, we present an extension of the FH model by considering the MoE framework, in light of the results obtained in the exploratory analysis of Section 3.1. To facilitate the presentation of the methodology and given our motivating application, we discuss an MoE constituted by two components. Specifically, we assume that the sampling level remains the same as (3), while an MoE is assumed at the linking level in place of (4), defining the proposed model that we label as FH-MoE. In other words, we assume that the latent area-level parameters are generated by a composite data-generating process, as a result of the intrinsic nature of the phenomenon under study. Conversely, we assume that the sampling distribution of the direct estimators is unaffected. In this spirit, we can express the FH-MoE linking level as follows:

$$\theta_d \mid \beta_{01}, \beta_{02}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_{u1}, \sigma_{u2}, z_d \overset{ind}{\sim} \mathbf{1}(z_d = 1)\mathcal{N}\left(\beta_{01} + \mathbf{x}_d^T\boldsymbol{\beta}_1, \sigma_{u1}^2\right) + \mathbf{1}(z_d = 2)\mathcal{N}\left(\beta_{02} + \mathbf{x}_d^T\boldsymbol{\beta}_2, \sigma_{u2}^2\right), \quad \forall d; \tag{7}$$

where z_d is a dichotomous latent variable that determines the cluster label for area d, assuming values 1 and 2. In addition, $\mathbf{1}(z_d = k)$ is an indicator variable that assumes value 1 if $z_d = k$ and zero otherwise, and each model component k is characterized by distinct coefficients $[\beta_{0k}, \boldsymbol{\beta}_k^T]^T$ and scale σ_{uk}.

The probability that z_d assumes value 1 is denoted as $\mathbb{P}[z_d = 1 \mid \mathbf{x}_d^*] = \pi(\mathbf{x}_d^*)$, where $\mathbf{x}_d^* \in \mathbb{R}^{P^*}$ is a vector of covariates possibly different to $\mathbf{x}_d$ used in modelling the location. In the MoE framework, this quantity is also defined as the gating network and can be interpreted as the prior probability that area d is assigned to the first mixture component. Specifically, for $\pi(\mathbf{x}_d^*)$, we assume logistic regression:

$$\log\left(\frac{\pi(\mathbf{x}_d^*)}{1-\pi(\mathbf{x}_d^*)}\right)\Bigg| \gamma_0, \boldsymbol{\gamma}, \upsilon_d = \gamma_0 + \mathbf{x}_d^{*T}\boldsymbol{\gamma} + \upsilon_d, \quad \forall d; \tag{8}$$

where γ_0 is the intercept, $\boldsymbol{\gamma} \in \mathbb{R}^{P*}$ is the vector of regression coefficients and $\upsilon_d \mid \sigma_\upsilon \overset{ind}{\sim} \mathcal{N}(0, \sigma_\upsilon^2), \ \forall d$, is an area-specific random effect.

If the latent variable z_d is integrated out, the predictor for the parameter θ_d conditional on the random effects u_{d1} and u_{d2} is

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

$$\theta_d \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, u_{d1}, u_{d2}, \pi(\mathbf{x}_d^*) = \pi(\mathbf{x}_d^*)(\beta_{01} + \mathbf{x}_d^T\boldsymbol{\beta}_1 + u_{d1}) + [1 - \pi(\mathbf{x}_d^*)](\beta_{02} + \mathbf{x}_d^T\boldsymbol{\beta}_2 + u_{d2}), \quad \forall d; \quad (9)$$

noting that the non-linearity of the relationship between θ_d and the covariates is manifest. In addition, from the previous expression, the combination of the two distinct zero mean random effects can be collected and denoted as $\tilde{u}_d = \pi(\mathbf{x}_d^*)u_{d1} + [1 - \pi(\mathbf{x}_d^*)]u_{d2}$. It can be interpreted as the overall random effect of area d, and its distribution depends on the two variance components σ_{u1}^2 and σ_{u2}^2 and the prior group probability $\pi(\mathbf{x}_d^*)$:

$$\tilde{u}_d \mid \pi(\mathbf{x}_d^*), \sigma_{u1}^2, \sigma_{u2}^2 \sim \pi(\mathbf{x}_d^*)\mathcal{N}(0, \sigma_{u1}^2) + [1 - \pi(\mathbf{x}_d^*)]\mathcal{N}(0, \sigma_{u2}^2), \quad (10)$$

from which it follows that the prior expectation of $\tilde{u}_d$ is zero, and its prior variance is $\pi(\mathbf{x}_d^*)\sigma_{u1}^2 + [1 - \pi(\mathbf{x}_d^*)]\sigma_{u2}^2$. To complete the model formulation, we apply the same prior distribution used for the FH model to the component-specific parameters. Likewise, we assume a half-normal prior for the random effect scale σ_v. In addition, independent Gaussian distributions with a mean of zero and standard deviation of 2.5 are set for the coefficients in Equation (8).

It is important to remark that we consider a mixture of two components due to the motivating application and the tendency of small area models to be parsimonious. However, the results presented here can be readily extended to a model with multiple components, if needed. The proposed model can be contextualized in relation to certain area-level modelling approaches that use a finite mixture at the linking levels. For example, Chakraborty et al. (2016) proposed an FH model with the random effect as in Equation (10). However, in their case, they set a constant probability $\pi(\mathbf{x}_d^*) = \pi$ for all areas, focusing on handling outliers. Meanwhile, the K-dimensional finite mixture model proposed by Gershunskaya and Savitsky (2020) tackles the issue of deviation from linearity by imposing a shared variance for random effects while allowing for distinct intercept terms. Similar to the previously mentioned approach, the probabilities associated with the mixture components remain constant across the areas within this framework.

4.2.1 *About shrinkage process and design consistency*

To investigate the shrinkage process and design consistency of the proposed FH-MoE model, we study the posterior distribution of θ_d conditioned on the linking level parameters. To ease the notation, we store them into the vector $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1^T, \boldsymbol{\Psi}_2^T)$, where $\boldsymbol{\Psi}_k = (\beta_{0k}, \boldsymbol{\beta}_k, \sigma_{uk})$, for $k = 1, 2$. The main results can be summarized as follows.

Proposition 1 The posterior conditional expectation and variance of θ_d are expressed as

$$\mathbb{E}\left[\theta_d \mid \boldsymbol{\Psi}, \pi(\mathbf{x}_d^*), \hat{\bar{Y}}_d\right] = \hat{\bar{Y}}_d + \\ - \left[\pi_P(\mathbf{x}_d^*)\eta_{d1}(\hat{\bar{Y}}_d - \beta_{01} - \mathbf{x}_d^T\boldsymbol{\beta}_1) + [1 - \pi_P(\mathbf{x}_d^*)]\eta_{d2}(\hat{\bar{Y}}_d - \beta_{02} - \mathbf{x}_d^T\boldsymbol{\beta}_2)\right]; \quad (11)$$

$$\mathbb{V}\left[\theta_d \mid \boldsymbol{\Psi}, \pi(\mathbf{x}_d^*), \hat{\bar{Y}}_d\right] = \pi_P(\mathbf{x}_d^*)\mathbb{V}\left[\theta_d \mid \boldsymbol{\Psi}_1, z_d = 1, \hat{\bar{Y}}_d\right] + [1 - \pi_P(\mathbf{x}_d^*)]\times \\ \times \mathbb{V}\left[\theta_d \mid \boldsymbol{\Psi}_2, z_d = 2, \hat{\bar{Y}}_d\right] + \pi_P(\mathbf{x}_d^*)[1 - \pi_P(\mathbf{x}_d^*)]\times \\ \times \left\{\mathbb{E}\left[\theta_d \mid \boldsymbol{\Psi}_1, z_d = 1, \hat{\bar{Y}}_d\right] - \mathbb{E}\left[\theta_d \mid \boldsymbol{\Psi}_2, z_d = 2, \hat{\bar{Y}}_d\right]\right\}^2, \quad (12)$$

where $\eta_{dk} = S_d^2/(\sigma_{uk}^2 + S_d^2)$ is the shrinkage factor of the kth component, and $\pi_P(\mathbf{x}_d^*) = \mathbb{P}[z_s = 1 \mid \boldsymbol{\Psi}_k, \pi(\mathbf{x}_d^*), \hat{\bar{Y}}_d]$ is the posterior probability of belonging

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

to the first mixture component. Its expression is

$$\pi_P(\mathbf{x}_d^*) = \frac{\pi(\mathbf{x}_d^*) f_{\mathcal{N}}\left(\hat{\bar{Y}}_d; \beta_{01} + \mathbf{x}_d^T \boldsymbol{\beta}_1, \sigma_{u1}^2 + S_d^2\right)}{\pi(\mathbf{x}_d^*) f_{\mathcal{N}}\left(\hat{\bar{Y}}_d; \beta_{01} + \mathbf{x}_d^T \boldsymbol{\beta}_1, \sigma_{u1}^2 + S_d^2\right) + \left[1 - \pi(\mathbf{x}_d^*)\right] f_{\mathcal{N}}\left(\hat{\bar{Y}}_d; \beta_{02} + \mathbf{x}_d^T \boldsymbol{\beta}_2, \sigma_{u2}^2 + S_d^2\right)}.$$

Proof. See Section S1 of the online supplementary material. □

The posterior conditional expectation in (11) can be re-expressed by highlighting the role of the overall shrinkage factor

$$\tilde{\eta}_d = \pi_P(\mathbf{x}_d^*)\eta_{d1} + \left[1 - \pi_P(\mathbf{x}_d^*)\right]\eta_{d2} \tag{13}$$

as

$$\mathbb{E}\left[\theta_d \mid \boldsymbol{\Psi}, \pi(\mathbf{x}_d^*), \hat{\bar{Y}}_d\right] = (1 - \tilde{\eta}_d)\hat{\bar{Y}}_d + \pi_P(\mathbf{x}_d^*)\eta_{d1}(\beta_{01} + \mathbf{x}_d^T\boldsymbol{\beta}_1) + \left(1 - \pi_P(\mathbf{x}_d^*)\right)(1 - \eta_{d2})(\beta_{02} + \mathbf{x}_d^T\boldsymbol{\beta}_2).$$

Furthermore, whenever $\beta_{01} = \beta_{02} = \beta_0$ $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}$, a special case in which only variance components are assumed to be group specific, (11) reduces to $\mathbb{E}[\theta_d \mid \boldsymbol{\Psi}, \pi(\mathbf{x}_d^*), \hat{\bar{Y}}_d] = (1 - \tilde{\eta}_d)\hat{\bar{Y}}_d + \tilde{\eta}_d(\beta_0 + \mathbf{x}_d^T\boldsymbol{\beta})$, a structure that readily reminds of the weighted average structure of the popular FH predictor in (6).

The results provide valuable insights into the model behaviour. For example, when an area is almost certain to be assigned to a specific group, i.e. when $\pi_P(\mathbf{x}_d^*) \to 0$ or $\pi_P(\mathbf{x}_d^*) \to 1$, the posterior expectation and variance converge to those associated with the component to which the area belongs. This fact highlights the model's ability to capture and reflect the characteristics of the dominant group when the membership becomes highly certain. Conversely, when the assignment of an area to a model component is uncertain, the posterior results stem from a mixture of the components, leading to variance inflation, as seen in the latter term of Equation (12). This seems a reasonable source of additional variability, particularly when the expectations of the two components are distant from each other.

Finally, we focus on the design consistency of the predictor $\mathbb{E}[\theta_d \mid \boldsymbol{\Psi}, \pi(\mathbf{x}_d^*), \hat{\bar{Y}}_d]$ of θ_d under our model. This property guarantees that as the area-specific sample size grows, the model-based predictor converges to the survey-weighted estimator, which is, in turn, consistent under the probability distribution induced by the survey design. In this sense, it offers, at least for areas with an adequate area-specific sample size, protection against model failure. The result can be formally stated as follows.

Proposition 2 In the asymptotic framework for finite population described by Isaki and Fuller (1982), the predictor (11) is design consistent, namely

$$\mathbb{E}\left[\theta_d \mid \boldsymbol{\Psi}, \pi(\mathbf{x}_d^*), \hat{\bar{Y}}_d\right] \xrightarrow{p} \bar{Y}_d, \tag{14}$$

where the convergence is in probability with respect to the randomization induced by sampling design.

Proof. As n_d grows large, the $S_d \longrightarrow 0$ and so do η_{d1}, η_{d2}. This implies that $\mathbb{E}[\theta_d \mid \boldsymbol{\Psi}, \pi(\mathbf{x}_d^*)\hat{\bar{Y}}_d] \xrightarrow{p} \hat{\bar{Y}}_d$, which, according to the definition given in (1), is a design consistent estimator. □

4.3 Posterior inference

The posterior inference was performed using Markov Chain Monte Carlo (MCMC) methods. Specifically, the implementation involved using the Hamiltonian Monte Carlo (HMC)

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

sampling algorithm through the `Stan` language and `rstan` package (Carpenter et al., 2017). The estimation was performed using four parallel chains, each comprising 6,000 iterations, where the initial 3,000 iterations were designated as warm-up and subsequently discarded. The Gelman–Rubin statistic and effective sample sizes are monitored to check for convergence or autocorrelation issues. Concerning the FH-MoE model, label-switching problems are avoided since cluster-specific intercepts β_{01} and β_{02} are constrained to be ordered, namely $\beta_{01} < \beta_{02}$. In any case, this issue would not affect θ_d because the linear combination of the two components retains its significance.

A crucial aspect to investigate when dealing with Bayesian mixture models is identifiability. Following the indications of Gormley and Frühwirth-Schnatter (2019, Section 5), we note that a careful assessment of the MCMC algorithm is crucial to verify that the model is identified by the data. In addition, when model complexity increases (e.g. more covariates, groups and presence of structured random effects), basic HMC algorithms implemented via `Stan` can experience difficulties. A possible solution to this problem is the implementation of algorithms that combine Gibbs steps with data augmentation strategies, useful to efficiently sample from a multinomial distribution (Gormley & Frühwirth-Schnatter, 2019; Polson et al., 2013; Zens et al., 2024).

Point estimators, designated as model-based or hierarchical Bayes (HB) estimators, are derived from the posterior expectations of $\theta_d \mid \text{data}$. Let us denote them as $\hat{\theta}_d = \mathbb{E}[\theta_d \mid \text{data}]$, whose evaluation, along with other posterior summaries, such as credible intervals and posterior variances, can be achieved by leveraging MCMC draws. The posterior quantities related to OOS areas, denoted as $\theta_{dj}^{\text{OOS}} \mid \text{data}$, can be retrieved by exploiting predictor equations, namely (5) for FH and (9) for FH-MoE. Specifically, areal covariates are combined with the posterior draws of parameters upon which predictors depend. Concerning area-specific parameters from unobserved areas, such as random effects, we propagate the uncertainty by drawing samples from their prior distributions. To better explain this, let us use a generic notation $\tilde{\cdot}_d^{(b)}$ to denote a posterior draw at the bth Monte Carlo replicate. For the FH-MoE model, the probability of an OOS area d to pertain to cluster 1, namely $\pi^{\text{OOS}}(\mathbf{x}_d^*)$, at the bth Monte Carlo replicate is obtained as follows:

$$\tilde{\pi}^{\text{OOS}}(\mathbf{x}_d^*)^{(b)} = \text{expit}\left(\tilde{\gamma}_0^{(b)} + \mathbf{x}_d^{*T}\tilde{\boldsymbol{\gamma}}^{(b)} + \tilde{v}_d^{(b)}\right), \quad \forall d; \tag{15}$$

in line with Equation (8), with the area-specific parameter $\tilde{v}_d^{(b)}$ generated from the prior $v_d \mid \tilde{\sigma}_v^{(b)} \sim \mathcal{N}(0, \tilde{\sigma}_v^{(b)})$. Locations of the different clusters are generated for OOS areas by adopting the same strategy and all those quantities are combined to retrieve $\theta_d^{\text{OOS}} \mid \text{data}$ as in (9). This can be trivially generalized for the FH model.

5 Simulation study

In this section, we discuss a simulation study aimed at evaluating the frequentist properties of the HB estimators for θ_d derived from various area-level models. Specifically, we compare the model-based estimates under the MoE-FH model, as introduced in Section 4.2, with alternative model specifications and the well-known FH model. The simulation study is model-based, scrutinizing the behaviour of these models under diverse data-generating processes.

The parameters governing the simulation exercise are specified to mimic our application. We consider a total of $D = 200$ areas, incorporating a subset of $D_{\text{OOS}} = 70$ OOS areas, to test predictors under such conditions. Direct estimates are generated at each replicate, with a total of $B = 1{,}000$ Monte Carlo replicates. To do so, the rule (3) is always adopted at the sampling level, whereas different scenarios are considered for θ_d at the linking level. The sampling variances S_d^2 are drawn following the rule: $S_d/2 \sim \text{Beta}(1.5, 3)$. A unique covariate x_d is generated, following two distinct distributions: 150 areas are drawn from $\mathcal{N}(0, 0.8^2)$, and 50 areas from $\mathcal{N}(3, 0.7^2)$; subsequently, the resulting values are standardized. These quantities are generated once and kept constant across the Monte Carlo iterations of the simulation study.

The five scenarios used to obtain θ_d are listed below.

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

- *Scenario 1*. To mimic a situation where the basic FH model holds, we set $\theta_d \sim \mathcal{N}(-0.1 + 0.7x_d, 0.2^2)$.
- *Scenario 2*. Then, we consider a situation in which the population is partitioned into two distinct clusters with different locations and variances. Here, the cluster assignment probability does not depend on the covariate:

$$\theta_d \,|\, z_d \sim \mathbf{1}(z_d = 1)\mathcal{N}\left(-0.7 + 0.5x_d, 0.1^2\right) + \mathbf{1}(z_d = 2)\mathcal{N}\left(0.7, 0.5^2\right),$$

  where $\mathbb{P}[z_d = 1] = 0.7$.

In the subsequent scenarios, the probability of belonging to the first cluster, i.e. $\mathbb{P}[z_d = 1] = \pi(x_d)$, depends on covariate x_d by relying on the logit transformation:

$$\log\left(\frac{\pi(x_d)}{1 - \pi(x_d)}\right) = -0.3 + 2.7x_d + \upsilon_d,$$

where $\upsilon_d \sim \mathcal{N}(0, 0.2^2)$. Then, the following three scenarios are set.

- *Scenario 3*. Two clusters having equal locations but different variances:

$$\theta_d \,|\, z_d \sim \mathbf{1}(z_d = 1)\mathcal{N}\left(-0.1 + 0.7x_d, 0.1^2\right) + \mathbf{1}(z_d = 2)\mathcal{N}\left(-0.1 + 0.7x_d, 0.5^2\right).$$

- *Scenario 4*. Two clusters having different locations but equal variances:

$$\theta_d \,|\, z_d \sim \mathbf{1}(z_d = 1)\mathcal{N}\left(-0.7 + 0.5x_d, 0.2^2\right) + \mathbf{1}(z_d = 2)\mathcal{N}\left(0.7, 0.2^2\right).$$

- *Scenario 5*. Two clusters having different locations and different variances:

$$\theta_d \,|\, z_d \sim \mathbf{1}(z_d = 1)\mathcal{N}\left(-0.7 + 0.5x_d, 0.1^2\right) + \mathbf{1}(z_d = 2)\mathcal{N}\left(0.7, 0.5^2\right).$$

In each scenario, four model specifications are evaluated. Each model relies on the sampling level of Equation (3), whereas different settings are specified for the linking level. The first considered model is the standard Bayesian FH model, with the linking level as in (4). Secondly, we set a mixture model (FH-Mix) with constant probability π and the following linking level:

$$\theta_d \,|\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_{u1}, \sigma_{u2}, \pi \sim \pi\mathcal{N}\left(\beta_{01} + x_d^T\beta_{11}, \sigma_{u1}^2\right) + [1 - \pi]\mathcal{N}\left(\beta_{02} + x_d^T\beta_{12}, \sigma_{u2}^2\right).$$

Then, we move to models involving an MoE specification. Firstly, a model with components having the same coefficients but different variances (FH-MoE-V) is specified:

$$\theta_d \,|\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_{u1}, \sigma_{u2}, \pi(x_d) \sim \pi(x_d)\mathcal{N}\left(\beta_0 + x_d^T\beta_1, \sigma_{u1}^2\right) + [1 - \pi(x_d)]\mathcal{N}\left(\beta_0 + x_d^T\beta_1, \sigma_{u2}^2\right).$$

Finally, the proposed FH-MoE model as in Equation (7) is considered. The prior distributions for the model parameters are as described in Section 4, while a uniform prior on the [0, 1] interval is assumed for the parameter π in the FH-Mix model.

Let θ_d represent the true value for area d and $\hat{\theta}_d^{(b)}$ denote the corresponding estimate at iteration b under the considered models. In determining the 90% credible intervals, we examine the 5th and 95th posterior percentiles, identified as $\theta_{d,L}^{(b)}$ and $\theta_{d,U}^{(b)}$ respectively. We define bias, root mean squared error (RMSE), coverage for 90% credible intervals, and the associated width as

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

$$\text{Bias}[\hat{\theta}_d] = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_d^{(b)} - \theta_d\right), \quad \text{RMSE}[\hat{\theta}_d] = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_d^{(b)} - \theta_d\right)^2},$$

$$\text{Coverage}[\theta_d] = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\left\{\theta_d \in \left[\theta_{d,L}^{(b)}; \theta_{d,U}^{(b)}\right]\right\}, \quad \text{Width}[\theta_d] = \frac{1}{B}\sum_{b=1}^{B}\left(\theta_{d,U}^{(b)} - \theta_{d,L}^{(b)}\right).$$

As summary metrics, these area-specific indicators are averaged across all areas (with the absolute value also taken for bias), distinguishing between IS and OOS areas. Moreover, for scenarios involving two distinct clusters, these summaries are computed separately for each group. The leave-one-out cross-validation information criterion (LOOIC, Vehtari et al., 2017) is monitored for model comparison purposes.

The summary outcomes from the simulation exercise are presented in Table 2 (average RMSEs), Table S1 of the online supplementary material (average bias and average absolute bias), Table 3 (average frequentist coverage of 90% credible intervals), Table 4 (average widths of credible intervals), and Table 5 (average LOOIC). Upon reviewing the results, initial observations regarding Scenario 1, i.e. without the existence of two distinct groups, can be made. The performances of the four models considered in the simulation study appear similar, suggesting that adopting mixture models in this context does not result in a significant deterioration of estimator performance. Another general remark can be made by comparing the performances of the FH and FH-Mix models. The results exhibit substantial overlap, indicating that fixing the prior probability that the areas belong to the first mixture component yields outcomes very similar to a single-component model, in the considered scenarios. This is because the two groups are characterized by differences in location and varying group membership probabilities. In this context, the FH-Mix model fails to differentiate between the two sub-groups, reducing its effectiveness to that of the FH model.

Focusing on the average RMSE (Table 2), FH-MoE exhibits lower values for this indicator, except for Scenario 1 and the OOS areas of Scenarios 2 and 3, where a slight overfitting issue could be hypothesized. Particularly noteworthy is the gain in accuracy observed when the two groups are characterized by different location rules (Scenarios 4 and 5), showcasing an improvement compared with other model-based methods, especially in OOS areas. We also observe that RMSEs related to areas belonging to Group 2 are higher due to the data-generating process. Indeed, such areas are characterized by a higher level of variability (in Scenarios 2, 3, and 5) or the covariate does not provide information on their location (Scenarios 2, 4, and 5).

In analysing the average bias and average absolute bias results presented in online supplementary material Table S1, several remarkable observations are worthy. Notably, the models exhibit comparable performance overall, with a slightly elevated absolute bias observed for FH-MoE. A compelling exception occurs in Scenarios 4 and 5, where the FH-MoE significantly outperforms its counterparts, underscoring its efficacy in these specific data-generating processes. The overall absolute bias values of the FH-MoE model, although slightly higher in some cases, generally result from the aggregation of similar values obtained for areas belonging to Groups 1 and 2. This implies that a model capable of identifying the two distinct sub-populations, such as the FH-MoE and FH-MoE-V models, does not distinctly penalize one of the two groups. Indeed, we can note that the FH and FH-Mix models show clearly unbalanced performances in the two groups. This is even more clear when moving to the average bias results of Table S1 of the online supplementary material. In scenarios characterized by clusters with different locations (2, 4, and 5), the FH, FH-Mix, and FH-MoE-V models tend to overestimate in areas belonging to the Group 1 and highly underestimate in areas of Group 2, essentially averaging everything out. The FH-MoE model, instead, mitigates this trend with improved and comparable bias performance across clusters.

A similar pattern can be observed for the average frequentist coverages reported in Table 3. Focusing on IS areas, values near the nominal level of 0.9 are obtained, noting that the FH and FH-Mix models are characterized by a slight over-coverage in Group 1 and under-coverage in Group 2, especially in Scenarios 3 and 5. This trend becomes even more evident in OOS areas: the FH and FH-Mix models produce intervals that over-cover the areas in Group 1 (almost always assuming value 1), and strongly under-cover areas in Group 2. Therefore, even if the overall values

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

Table 2. Average RMSE distinguishing between IS and OOS areas

		IS areas			OOS areas		
Scen.	Model	Group 1	Group 2	Overall	Group 1	Group 2	Overall
1	FH	–	–	0.106	–	–	0.176
	FH-Mix	–	–	0.106	–	–	0.176
	FH-MoE-V	–	–	0.107	–	–	0.176
	FH-MoE	–	–	0.109	–	–	0.182
2	FH	0.158	0.158	0.158	0.542	1.032	0.717
	FH-Mix	0.158	0.158	0.158	0.542	1.032	0.717
	FH-MoE-V	0.140	0.160	0.147	0.413	1.239	0.708
	FH-MoE	0.119	0.158	0.133	0.528	1.099	0.732
3	FH	0.111	0.148	0.123	0.088	0.439	0.223
	FH-Mix	0.112	0.148	0.123	0.088	0.439	0.223
	FH-MoE-V	0.085	0.142	0.104	0.093	0.435	0.225
	FH-MoE	0.077	0.142	0.098	0.104	0.441	0.234
4	FH	0.131	0.150	0.137	0.186	0.494	0.305
	FH-Mix	0.131	0.149	0.137	0.186	0.494	0.305
	FH-MoE-V	0.116	0.148	0.126	0.179	0.509	0.306
	FH-MoE	0.121	0.123	0.122	0.201	0.361	0.263
5	FH	0.130	0.159	0.139	0.120	0.625	0.315
	FH-Mix	0.130	0.158	0.139	0.120	0.625	0.315
	FH-MoE-V	0.098	0.155	0.117	0.101	0.682	0.325
	FH-MoE	0.080	0.148	0.102	0.112	0.612	0.305

Note. Distinct values are reported for areas belonging to Groups 1 and 2.

Table 3. Average frequentist coverage levels (related to 90% credible intervals) computed distinguishing between IS and OOS areas

		IS Areas			OOS Areas		
Scen.	Model	Group 1	Group 2	Overall	Group 1	Group 2	Overall
1	FH	–	–	0.906	–	–	0.866
	FH-Mix	–	–	0.907	–	–	0.868
	FH-MoE-V	–	–	0.905	–	–	0.869
	FH-MoE	–	–	0.878	–	–	0.8522
2	FH	0.905	0.893	0.901	1.000	0.757	0.913
	FH-Mix	0.904	0.893	0.900	1.000	0.755	0.913
	FH-MoE-V	0.914	0.885	0.904	1.000	0.613	0.862
	FH-MoE	0.925	0.894	0.914	0.945	0.682	0.851
3	FH	0.935	0.838	0.904	1.000	0.513	0.812
	FH-Mix	0.936	0.838	0.904	1.000	0.517	0.814
	FH-MoE-V	0.898	0.882	0.893	0.890	0.742	0.833
	FH-MoE	0.874	0.879	0.875	0.842	0.797	0.825
4	FH	0.916	0.860	0.898	0.980	0.688	0.867
	FH-Mix	0.916	0.862	0.899	0.979	0.688	0.867
	FH-MoE-V	0.904	0.887	0.899	0.936	0.730	0.856
	FH-MoE	0.877	0.878	0.877	0.868	0.776	0.833
5	FH	0.921	0.845	0.897	1.000	0.613	0.851
	FH-Mix	0.921	0.846	0.897	1.000	0.609	0.849
	FH-MoE-V	0.906	0.878	0.897	0.947	0.764	0.876
	FH-MoE	0.876	0.895	0.882	0.826	0.781	0.809

Note. Distinct values are reported for areas belonging to Groups 1 and 2.

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

Table 4. Average width (related to 90% credible intervals) computed distinguishing between IS and OOS areas

		IS areas			OOS areas		
Scen.	Model	Group 1	Group 2	Overall	Group 1	Group 2	Overall
1	FH	–	–	0.370	–	–	0.658
	FH-Mix	–	–	0.371	–	–	0.662
	FH-MoE-V	–	–	0.372	–	–	0.670
	FH-MoE	–	–	0.359	–	–	0.668
2	FH	0.527	0.507	0.520	2.549	2.544	2.547
	FH-Mix	0.527	0.507	0.520	2.539	2.535	2.538
	FH-MoE-V	0.483	0.506	0.492	2.607	2.651	2.623
	FH-MoE	0.418	0.507	0.450	2.259	2.262	2.260
3	FH	0.439	0.411	0.430	0.952	0.956	0.953
	FH-Mix	0.440	0.412	0.431	0.955	0.959	0.957
	FH-MoE-V	0.310	0.448	0.355	0.516	1.363	0.843
	FH-MoE	0.269	0.450	0.328	0.394	1.489	0.816
4	FH	0.473	0.443	0.464	1.222	1.228	1.224
	FH-Mix	0.474	0.443	0.464	1.225	1.230	1.226
	FH-MoE-V	0.413	0.475	0.433	0.854	1.731	1.192
	FH-MoE	0.396	0.390	0.394	0.871	1.038	0.935
5	FH	0.480	0.450	0.470	1.294	1.299	1.296
	FH-Mix	0.480	0.450	0.470	1.296	1.301	1.298
	FH-MoE-V	0.353	0.482	0.395	0.709	2.032	1.219
	FH-MoE	0.279	0.483	0.345	0.488	1.856	1.015

Note. Distinct values are reported for areas belonging to Groups 1 and 2.

Table 5. Average LOOIC obtained by different models across various scenarios

Scen.	FH	FH-Mix	FH-MoE-V	FH-MoE
1	−75.36	−76.51	−75.92	−75.11
2	−25.65	−25.11	−41.35	−62.22
3	−58.17	−59.57	−80.26	−84.91
4	−37.57	−38.02	−49.05	−62.27
5	−34.35	−34.23	−61.32	−79.57

of the FH-MoE model are often lower when aggregated over the whole set of areas, the intervals have similar coverage for all areas when stratifying with respect to the group, being more reliable than those obtained with the FH and FH-Mix models. Slightly higher coverage values are obtained under model FH-MoE-V, which tends to produce intervals with moderate over-coverage for Group 1.

This is even more interesting if coupled with the average widths of the intervals reported in Table 4. Indeed, models relying on the MoE framework can differentiate the width of the intervals across the groups, which are generally narrower for group 1 and wider for Group 2. In this way, the coverage level is reached with more precise intervals for Group 1, keeping a reasonable coverage also for Group 2. This is in line with the data generation process of Scenarios 3 and 5, where the clusters are characterized by different variances (0.1^2 and 0.5^2, respectively), not captured by the FH and FH-Mix models. A further in-depth analysis of average widths is enclosed in Figure S1 in the online supplementary material. In particular, the average width of areas versus the standard deviation of group membership probabilities for OOS areas in Scenarios 3, 4, and 5 (i.e. those with varying $\pi(\boldsymbol{x}_d)$) is depicted. Specifically, when considering Scenario 4, i.e. the case in which group variances are equal, results are as expected: the more the group membership is uncertain, the more the posterior variability of predictions increases. Conversely, in Scenarios 3 and 5, where group

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

Table 6. LOOIC results for the models included in the analysis

	FH	FH-Mix	2FH	FH-MoE-V	FH-MoE
LOOIC (S.E.)	186.6 (35.6)	180.4 (35.5)	178.0 (35.5)	161.0 (32.9)	134.3 (28.7)

variabilities are assumed as unequal and unbalanced, the trend reverses for Group 2. In particular, the decline in variability near $\pi(\boldsymbol{x}_d) = 0.5$ is largely due to the large unbalance between group variances that starts averaging. This is consistent with the conditional posterior variance of predictions of Equation (12).

Table 5 summarize the average LOOIC values obtained from the different models. In Scenario 1 the performances of the models are similar; meanwhile, in the other scenarios, the FH-MoE model shows lower LOOIC values. In Scenario 3, i.e. when the data-generating process imposes equal regression coefficients for both groups, the FH-MoE-V and FH-MoE models have similar performances.

6 Results from application

This section discusses a poverty mapping analysis in Bangladesh. The small area models discussed in Section 4 are applied to the data described in Sections 2.1 and 3. Specifically, direct estimates of per capita WI at upazila level, as described in Section 2, serve as response variables. Such estimates are obtained in $D_{\text{IS}} = 365$ areas in which at least one sampled unit is available while producing reliable estimates for the $D_{\text{OOS}} = 179$ unsampled areas. In addition, the variances at the sampling levels are assumed to be known and equal to the complex survey variances obtained in Section 2, i.e. $S_d^2 = \mathbb{V}_{cs}[\hat{\bar{Y}}_d]$.

As auxiliary information, the RS covariates introduced in Section 3 and summarized in Table 1 are considered. However, we detected a multicollinearity issue, leading us to resort to the PCA performed in Section 3.1 to obtain uncorrelated regressors. Notably, a small area model does not have causal inference ambitions; thus, loss of interpretability may not be a concern. Moreover, such a procedure has already been exploited in SAE (e.g. see Brakel et al., 2016). In particular, we include the first two principal components as covariates both in the expert network of Equation (7) and in the gating network of Equation (8), so that $\mathbf{x}_d = \mathbf{x}_d^*, \forall d$. A detailed assessment of the convergence of the MCMC algorithm is discussed in Section S3.1 of the online supplementary material, while further details about the model selection process are presented in Section S3.2 of the online supplementary material. In particular, models containing spatial information, an additional cluster and more covariates are tested. None of these model specifications significantly improve the ones we discuss here.

The results obtained using the proposed FH-MoE model are first compared with those obtained using FH, FH-Mix, and FH-MoE-V models. As an additional model for comparison, we consider the separate estimation of the FH model on groups of areas defined by the explorative clustering performed in Section 3.1, denoted by 2FH. As a model selection tool, Table 6 reports the LOOIC values, from which it is possible to deduce that the predictive abilities of the FH-MoE model are superior to those of the more standard proposals. Indeed, its more flexible specification shows the lowest LOOIC value, confirming the simulation performances presented in Section 5. In light of these results and for clarity, we now restrict our attention to the results related to the FH and FH-MoE models.

Let us investigate the implicit classification performed by the FH-MoE model by comparing it with the explorative clustering detailed in Section 3.1. Because the classification relies on a fuzzy measure such as $\pi(\mathbf{x}_d)$, we establish the optimal cut-off (equal to 0.14) by minimizing the area under the ROC curve. Through discretization, we can readily compare the clusters identified by the two procedures, revealing a high degree of similarity, with only 2 of 544 areas having conflicting classifications. Thus, the interpretation of the two clusters remains very similar to the one described in Section 3.1, with one cluster including peri-urban, rural, and remote areas and the second one, smaller in size, encompassing metropolitan, and highly urbanized upazilas. Insights into the interpretation of groups obtained under the mixture models can also be gleaned from

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

Table 7. Posterior means and 90% credible intervals of the probability of belonging to cluster 1 (for the FH-Mix model) and the regression coefficients involved in the logit regression of Equation (8), related to the FH-MoE model

Model	Parameter	Est	90% CI
FH-MoE	γ_0	−0.26	[−1.30,1.09]
	γ_1	−2.85	[−4.67,−1.09]
	γ_2	−0.06	[−0.55,0.39]

Table 8. Posterior means and 90% credible intervals related to the basic parameters that govern the models at the linking levels

Comp.	FH		FH-MoE			
	Unique		Group 1		Group 2	
Par.	Est	90% CI	Est	90% CI	Est	90% CI
β_0	−0.08	[−0.11,−0.04]	−0.54	[−0.82,−0.32]	0.49	[0.20,0.99]
β_1	0.66	[0.62,0.70]	0.42	[0.10,0.73]	0.42	[0.23,0.55]
β_2	0.07	[0.04,0.11]	0.01	[−0.05,0.07]	0.13	[−0.07,0.32]
σ_u	0.33	[0.30,0.36]	0.11	[0.03,0.17]	0.52	[0.41,0.69]

Table 7, which presents posterior results related to the parameters governing the probability of an area belonging to the first component. Indeed, under the FH-MoE model, the probability is modelled according to the latent logit regression in (8). We can note that high scores of the first principal component strongly increase the probability of belonging to the second cluster (posterior mean of γ_1 equal to −2.85). This adds evidence to the fact that the first principal component highly identifies the two clusters as noted in the exploratory analysis (see Figure 4).

The outcomes pertaining to the basic model parameters at the linking levels, specifically regression coefficients and variance components, are detailed in Table 8. The FH-MoE model yields distinct estimates of parameters in its two components and also with respect to the FH model ones. Firstly, the posterior distributions of the intercepts are sharply separated, suggesting a substantial link of cluster 2 to areas with high WI values. Moreover, the second principal component appears to hold greater relevance in cluster 2. This makes sense because such a component takes greater values for city-centre upazilas, characterized by wealthier households. Meanwhile, the posterior means of coefficients related to the first principal component are equal, possibly because its impact is mostly absorbed by cluster-specific intercepts. Notably, the sign of the coefficients indicates that the higher the score of the first principal component, the higher the expected per capita WI; the posterior uncertainty related to such estimates indicates more precision for cluster 2. Lastly, a significant difference emerges in the estimation of the random effect scale parameter between the FH-MoE and FH models: under the FH-MoE model, the scale of cluster 1 is notably lower than that of cluster 2, whereas, in the FH model, the scale parameter lies at an intermediate level. This is reasonable because the direct estimates associated with urban areas are more accurate. Consequently, substantial effort is required for random effects to minimize shrinkage, aiming to achieve model-based estimates that closely align with the direct ones. When fitting the FH-Mix model, the results indicate that the two clusters are nearly equiprobable (with an estimated π equal to 0.49). In addition, the posterior means of the coefficients closely resemble those obtained under the FH model, suggesting that the model is ineffective in capturing the rural/urban divide.

The comparison between the proposed FH-MoE model and the basic FH model is enhanced by considering some estimated area-specific quantities, illustrated through scatterplots in Figure 5. In the left-side panel, the posterior mean of the area-specific random effect u_d in the FH model is juxtaposed with the posterior mean of $\tilde{u}_d$ from (10), representing the corresponding area-specific random effect in the mixture model. The central panel displays the posterior means of the shrinkage

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

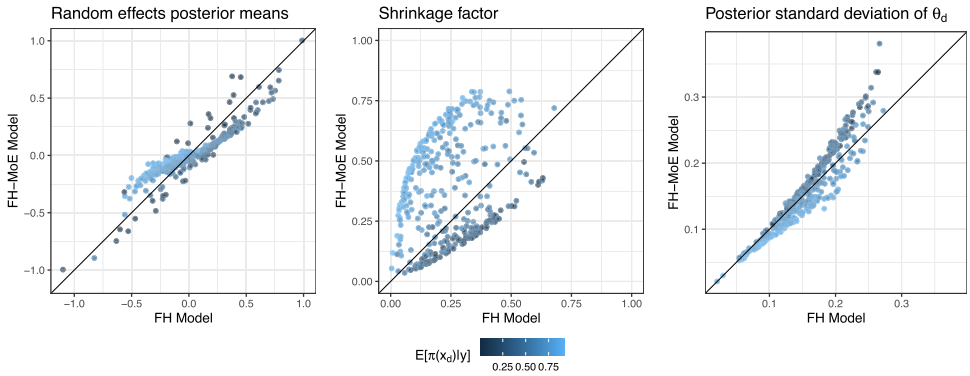


Figure 5. Scatterplots of area-specific quantities estimated under the FH and FH-MoE models: posterior means of random effects (left panel), shrinkage factor (central panel), and posterior standard deviation of θ_d (right panel). The colour scale depends on the posterior mean of $\pi(\mathbf{x}_d)$, the black line represents the bisector.

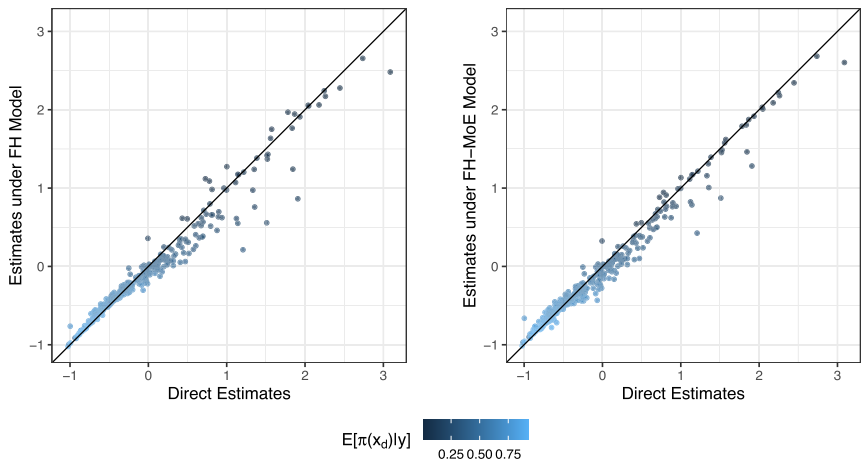


Figure 6. Scatterplots of direct estimates versus the model-based estimates under the FH and FH-MoE models. The colour scale depends on the posterior mean of $\pi(\mathbf{x}_d)$; the black line represents the bisector.

factors, where the definition (13) is used for the FH-MoE model. Finally, the right-hand side plot compares the posterior standard deviations obtained for θ_d. To aid interpretation, dots are colour-coded on a gradient scale reflecting the posterior mean of probability $\pi(\mathbf{x}_d)$. Light-blue dots denote areas likely belonging to Cluster 1, while dark-blue dots indicate areas associated with Group 2, characterized by higher per capita WI. In general, the FH-MoE model enables the capture of two distinct shrinkage processes. Notably, areas likely associated with cluster 1 exhibit a more pronounced shrinkage effect, as evidenced by the concentration of posterior means of random effects around zero. Referencing Table 8, the FH-MoE model yields a posterior mean of 0.11 for σ_{u1} compared with 0.33 for σ_u in the FH model. This discrepancy becomes evident when examining shrinkage factors, revealing markedly higher values for such areas in the FH-MoE model. Conversely, regions with a high probability of belonging to cluster 2 display less-shrunk random effects and lower shrinkage factors, recalling that the posterior mean of σ_{u2} is 0.52. This behaviour is reflected in the posterior standard deviations of θ_d: a noticeable reduction in standard deviation is observed for areas in cluster 1, whereas higher values are recorded for areas in cluster 2.

These distinctive model characteristics become also apparent when examining the shrinkage processes of model-based estimates in contrast to direct estimates (Figure 6). In the estimates derived from the FH model (left panel), there is a noticeable higher shrinkage of estimates in the right

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

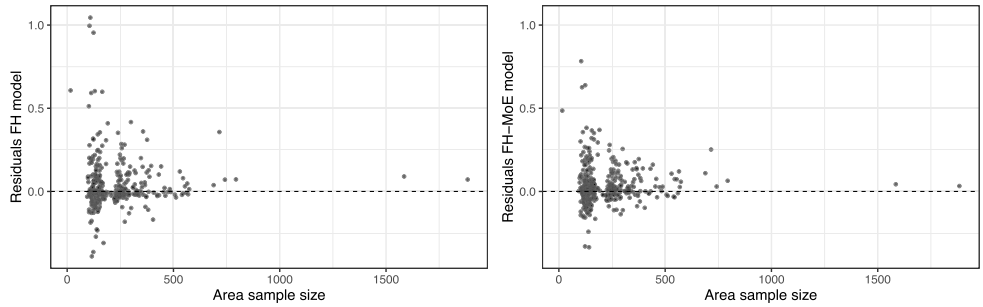


Figure 7. Scatterplots to assess the design consistency: residuals versus the sample size.

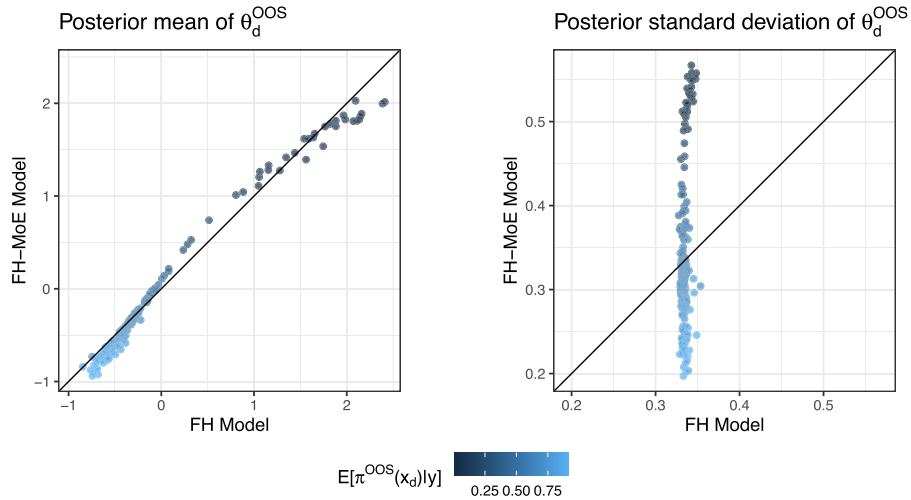


Figure 8. Scatterplots of posterior means and standard deviations of θ_d^{OOS}, comparing the FH and FH-MoE models. The colour scale depends on the posterior mean of $\pi^{\text{OOS}}(\mathbf{x}_d)$, the black line represents the bisector.

tail of the per capita WI distribution. Meanwhile, under the FH-MoE model (right panel), these areas experience less shrinkage towards the general mean, while estimates associated with areas in Cluster 1 appear to be significantly shrunk towards the group-specific mean. The substantial shrinkage induced by the FH model in regions with high per capita WI results in larger residuals than in the FH-MoE model. This observation is evident in Figure 7, where the design consistency property is confirmed for both the FH and FH-MoE models, as residuals approach zero with increasing sample size. Nevertheless, it appears that the residuals in domains with large sample sizes (i.e. $n_d > 1{,}000$) are closer to zero when using the FH-MoE model than when using the FH model.

Another interesting comparison between the FH and FH-MoE models regards the per capita WI estimates related to OOS areas. Figure 8 compares both the posterior means and standard deviations of θ_d^{OOS}, with a gradient colour scale determined by the predicted probability $\pi^{\text{OOS}}(\mathbf{x}_d)$. The presence of two regimes distinguished by FH-MoE (Figure 5), leads to notable differences in the estimates compared with those obtained using the standard FH model. From the posterior means (left panel), we can note the differences in the linear predictors that characterize the locations of the clusters, whereas the posterior standard deviations in the right panel show a marked heterogeneity for estimates using the FH-MoE model.

The latter result confirms the inadequacy of the FH model in capturing the uncertainty related to model-based estimates. The posterior standard deviation of the estimates is basically constant

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

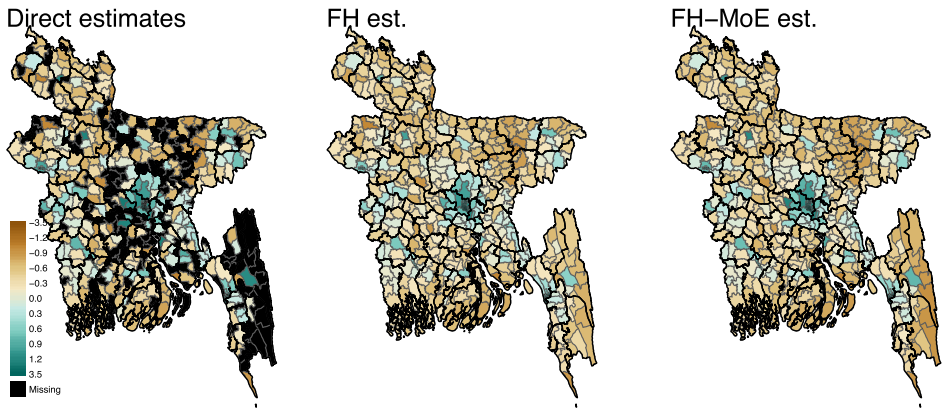


Figure 9. Maps of WI direct estimates (left-hand side plot), FH model-based estimates (central plot), and FH-MoE model-based estimates (right-hand side plot).

across areas, leading to overestimation for areas in cluster 1 and underestimation for areas in cluster 2. These results agree with those obtained in the simulation exercise concerning the clear imbalance of coverage of credible intervals between clusters and the constant interval widths (Tables 3 and 4).

Finally, a mapping of direct estimates, compared with model-based estimates from the FH and FH-MoE models, is set out in Figure 9. The map relating to direct estimates emphasizes the great amount of OOS areas. This problem can be solved by the predictive abilities of small-area models that also provide spatially smoothed estimates. The FH-MoE estimates, as compared with the FH estimates, are characterized by slightly higher WI values in predominantly urban areas, such as the Dhaka metropolitan area, due to the mild shrinking process. Lower WI values for the FH-MoE estimates than for the FH estimates, are registered for the remote (rural) areas on the Chittagong Hill Tract (south-east) consistent with Figure 8.

7 Concluding remarks

In this study, we propose an extension of the FH model by incorporating an MoE framework. This enhancement is particularly suitable when areas or domains exhibit heterogeneity due to factors or groupings. We applied it to the rural/urban divide related to the measurement of per capita WI in the Bangladeshi upazilas.

Our analysis revealed two distinct clusters: one rural cluster with lower per capita WI and a wealthier urban cluster with distinct relationships with RS covariates. This situation is likely to characterize many middle- and lower-income countries (see for instance Ejaz & Mallawaarachchi, 2023; Wu & He, 2018), making our application more general. In our specific application, the MoE methodology enables a more pronounced shrinkage in the rural cluster than in the urban cluster, addressing shortcomings of the standard FH model which would imply uniform shrinkage across areas with differing characteristics. Generally, the FH-MoE model may outperform the FH model whenever there is heterogeneity in regression coefficients and variance components, a situation that is likely to occur in many small area estimation problems (see the discussion in Lahiri & Salvati, 2023). In addition, synthetic components to which we shrink differ between clusters due to variations in the relationship between the auxiliary and target variables.

A simulation study identifies such an approach to be satisfactory in terms of improved performance of model-based estimators, compared with the FH model and some alternative specifications. Indeed, in scenarios with distinct sub-populations of areas, the proposed estimator yields balanced performance between distinct clusters. Conversely, the FH model tends to penalize one cluster with respect to the other in terms of bias, error, and coverage. This flexibility also improves

Downloaded from academic.oup.com/jrssc/article/74/5/1255/8096259 by guest on 28 September 2026

predictions for OOS areas as the process involves predicting gating probabilities followed by averaging predictions from different experts.

Our methodology can be extended in several ways, such as considering more than two experts using multinomial logistic regression in the gating step, if the analysed data require a more complex model. A more general specification in this line is to consider the number of clusters as unknown, placing a prior distribution on this parameter. In addition, each expert can be represented by a different model, allowing for variations in auxiliary information, distributional assumptions or dependency structures. This approach accommodates situations where each expert has a distinct specification, resulting in area-level predictions derived from a mixture of different estimators weighted by data-driven composition weights. Another useful model extension could be introducing a prior distribution aimed to selecting relevant regressors to avoid possible overfitting issues (Ghosh et al., 2011; Lu & Lou, 2023).

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.

Conflicts of interest: None declared.

Funding

The work of Silvia De Nicolò was partially funded by 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 Ministero dell'Università e della Ricerca on funds FSE REACT EU - PON R&I 2014–2020 and PNR (D.M. 737/2021) for the RTDA_GREEN project under Grant J41B21012140007.

Data availability

The DHS data used in the analysis is not available due to the data disclosure agreement. The code to run the FH and FH-MoE models through Stan with pseudo-data (one iteration of the model-based simulation of Section 5) is made available in the online supplementary material.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

Brakel J. A., Buelens B., & Boonstra H.-J. (2016). Small area estimation to quantify discontinuities in repeated sample surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *179*(1), 229–250. https://doi.org/10.1111/rssa.12110

Carpenter B., Gelman A., Hoffman M. D., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo J., Li P., & Riddell A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Casas-Cordero Valencia C., Encina J., & Lahiri P. (2016). Poverty mapping for the Chilean comunas. In *Analysis of Poverty Data by Small Area Estimation* (pp. 379–404). Wiley Online Library.

Chakraborty A., Datta G. S., & Mandal A. (2016). A two-component normal mixture alternative to the Fay-Herriot model. *Statistics in Transition new Series*, *17*(1), 67–90.https://doi.org/10.59170/stattrans-2016-004

Chi G., Fang H., Chatterjee S., & Blumenstock J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(3), e2113658119. https://doi.org/10.1073/pnas.2113658119

Corsi D. J., Neuman M., Finlay J. E., & Subramanian S. (2012). Demographic and health surveys: A profile. *International Journal of Epidemiology*, *41*(6), 1602–1613. https://doi.org/10.1093/ije/dys184

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026

Diz-Rosales N., Lombardía M. J., & Morales D. (2024). Poverty mapping under area-level random regression coefficient Poisson models. *Journal of Survey Statistics and Methodology*, *12*(2), 404–434. https://doi.org/10.1093/jssam/smad036

Ejaz N., & Mallawaarachchi T. (2023). Disparities in economic achievement across the rural–urban divide in Pakistan: Implications for development planning. *Economic Analysis and Policy*, *77*(7), 487–512. https://doi.org/10.1016/j.eap.2022.11.023

Fay R. E., & Herriot R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277. https://doi.org/10.1080/01621459.1979.10482505

Ferdousi S., & Dehai W. (2014). Economic growth, poverty and inequality trend in Bangladesh. *Asian Journal of Social Sciences & Humanities*, *3*(1), 1–11.

Frühwirth-Schnatter S., Celeux G., & Robert C. P. (2019). *Handbook of mixture analysis*. CRC Press.

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

Garrett J., & Chowdhury S. (2004). *Urban-rural links and transformation in Bangladesh: A review of the issues* (Technical Report), International Food Policy Research Institute (IFPRI).

Gershunskaya J., & Savitsky T. D. (2020). Model-based screening for robust estimation in the presence of deviations from linearity in small domain models. *Journal of Survey Statistics and Methodology*, *8*(2), 181–205. https://doi.org/10.1093/jssam/smz004

Ghosh J., Herring A. H., & Siega-Riz A. M. (2011). Bayesian variable selection for latent class models. *Biometrics*, *67*(3), 917–925. https://doi.org/10.1111/biom.2011.67.issue-3

Goodrich B., Gabry J., Ali I., & Brilleman S. (2023). rstanarm: Bayesian applied regression modeling via Stan, https://mc-stan.org/rstanarm/. R package version 2.21.4.

Gormley I. C., & Frühwirth-Schnatter S. (2019). Mixture of experts models. In *Handbook of mixture analysis* (pp. 271–307). Chapman and Hall/CRC.

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

Hobza T., & Morales D. (2013). Small area estimation under random regression coefficient models. *Journal of Statistical Computation and Simulation*, *83*(11), 2160–2177. https://doi.org/10.1080/00949655.2012.684094

Isaki C. T., & Fuller W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*(377), 89–96. https://doi.org/10.1080/01621459.1982.10477770

Kalton G. (1979). Ultimate cluster sampling. *Journal of the Royal Statistical Society: Series A (General)*, *142*(2), 210–222. https://doi.org/10.2307/2345081

Kish L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, *8*(2), 183.

Lahiri P., & Salvati N. (2023). A nested error regression model with high-dimensional parameter for small area estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(2), 212–239. https://doi.org/10.1093/jrsssb/qkac010

Lee K., & Braithwaite J. (2022). High-resolution poverty maps in sub-saharan Africa. *World Development*, *159*, 106028. https://doi.org/10.1016/j.worlddev.2022.106028

Liu B., Lahiri P., & Kalton G. (2014). Hierarchical Bayes modeling of survey-weighted small area proportions. *Survey Methodology*, *40*(1), 1–13.

Lu Z., & Lou W. (2023). Bayesian approaches to variable selection in mixture models with application to disease clustering. *Journal of Applied Statistics*, *50*(2), 387–407. https://doi.org/10.1080/02664763.2021.1994529

Lumley T. (2011). *Complex surveys: A guide to analysis using R*. John Wiley & Sons.

Maiti T., Ren H., Dass S. C., Lim C. Y., & Maier K. S. (2014). Clustering based small area estimation: An application to Meap data. *Calcutta Statistical Association Bulletin*, *66*(1-2), 73–94. https://doi.org/10.1177/0008068320140106

Nguyen H. D., & Chamroukhi F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/widm.1246

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.

Peng F., Jacobs R. A., & Tanner M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, *91*(435), 953–960. https://doi.org/10.1080/01621459.1996.10476965

Poirier M. J., Grépin K. A., & 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), 1–46. https://doi.org/10.1007/s11205-019-02187-9

Downloaded from academic.oup.com/jrssc/article/74/5/1255/8096259 by guest on 28 September 2026

Polettini S. (2017). A generalised semiparametric Bayesian Fay-Herriot model for small area estimation shrinking both means and variances. *Bayesian Analysis*, *12*(3), 729–752. https://doi.org/10.1214/16-BA1019

Polson N. G., Scott J. G., & Windle J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Prasad N. N., & Rao J. N. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*(409), 163–171. https://doi.org/10.1080/01621459.1990.10475320

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

Rigg J. (1998). Rural–urban interactions, agriculture and wealth: A southeast Asian perspective. *Progress in Human Geography*, *22*(4), 497–522. https://doi.org/10.1191/030913298667432980

Rutstein S. O., & Johnson K. (2004). *DHS comparative reports no. 6: The DHS Wealth Index* (Technical Report). ORC Macro, Calverton, Maryland, USA.

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., Hadiuzzaman KN, Lu X, Wetter E, Tatem AJ, & Bengtsson L (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society, Interface*, *14*(127), 20160690. https://doi.org/10.1098/rsif.2016.0690

Torkashvand E., Jafari Jozani M., & Torabi M. (2017). Clustering in small area estimation with area level linear mixed models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(4), 1253–1279. https://doi.org/10.1111/rssa.12308

Vehtari A., Gelman A., & Gabry J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wu J.-X., & He L.-Y. (2018). Urban–rural gap and poverty traps in China: A prefecture level analysis. *Applied Economics*, *50*(30), 3300–3314. https://doi.org/10.1080/00036846.2017.1420890

Yeh C., Perez A., Driscoll A., Azzari G., Tang Z., Lobell D., Ermon S., & Burke M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-019-13993-7

You Y., & Hidiroglou M. (2023). Application of sampling variance smoothing methods for small area proportion estimation. *Journal of Official Statistics*, *39*(4), 571–590. https://doi.org/10.2478/jos-2023-0026

Zens G., Frühwirth-Schnatter S., & Wagner H. (2024). Ultimate Pólya Gamma samplers—Efficient MCMC for possibly imbalanced binary and categorical data. *Journal of the American Statistical Association*, *119*(548), 2548–2559. https://doi.org/10.1080/01621459.2023.2259030

Downloaded from academic.oup.com/jrsssc/article/74/5/1255/8096259 by guest on 28 September 2026