Mapping Well-Being Through a Mixture-of-Experts Fay-Herriot Model

Mapping well-being through a Mixture-of-Experts Fay-Herriot model

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

[1] Department of Statistical Sciences, Università di Bologna,
aldo.gardini@unibo.it, silvia.denicolo@unibo.it
[2] DISES & DSS, Università Cattolica del S. Cuore
enrico.fabrizi@unicatt.it

Abstract. Our research is motivated by estimation of the per capita wealth index in Bangladeshi upazilas, integrating data from the Demographic and Health Survey along with remote sensing covariates, in a small area estimation framework. The popular Fay-Herriot model shows relevant limitations when applied to our data, as it fails to manage adequately two (or more) distinct regimes in the data generating process, such as those implied by the rural/urban divide that characterizes many low and middle income countries. We extend the Fay-Herriot model through a Mixture-of-Experts, where areas are classified into groups within the estimation process. This adds flexibility while keeping relevant properties of the predictors such as design consistency and easy interpretation. In addition, this family of models defines the mixing probabilities through a logistic regression, turning out to be particularly convenient in the applied setting.

Keywords: Bayesian inference; DHS survey; small area estimation;well-being indicators

1 Introduction

Estimates of socio-economic indicators are often needed for small subsets of a population in order to monitor their geographical and/or social distribution. The primary data source for estimation is often a sample survey, which typically fall short of providing reliable estimates for small sub-populations (small area) because of limited area-specific sample sizes. To improve the reliability of these estimates, small area estimation methods integrate survey data with other data sources. The different data sources can be integrated at the unit/household level or at the area of interest level as we do in this research (see [5], Ch. 6 and 7).

The Fay-Herriot (FH) [2] model is the simplest and most popular area-level model. One distinguishing feature of this model is the linear regression that links the area-level target parameter to auxiliary information. This entails the assumption of a single intercept and regression slopes for the whole ensemble of areas being considered. Such an assumption can be very restrictive when studying problems characterised by a relationship between response and covariates that varies across the areas.

The need for an extension of FH model is motivated by an application. Specifically, we consider the estimation of the per capita wealth index (WI) in Bangladeshi upazilas, i.e., small administrative subdivisions. Direct survey estimates are obtained from the Demographic and Health Survey (DHS), while auxiliary information comes mostly from remote sensing (RS) satellite images. A rural/urban divide in the WI distribution is typically observed for developing countries, induced by markedly different economies. From the statistical perspective, this can translate into the existence of separate data generating processes. This step could be nontrivial in area-level models as the rural/urban classification of territories does not perfectly overlap with administrative boundaries. As a consequence, the areas cannot be classified a priori.

In this paper, we face this challenge by proposing a Bayesian Mixture-of-Experts (MoE) model [3] that extends the FH model by allowing for the existence of different groups of areas each characterised by distinct predictors. An important feature of MoE is the modelling of the area-specific probability of belonging to a given group (known also as *gating* probabilities) through a logistic regression. Each model component (i.e. the *expert*) is a distinct Fay-Herriot model, namely allowing for different coefficients and/or area-specific random effects.

Although the framework we are dealing with relates to previous proposals in the small area literature (e.g., [6]), it presents distinct novelties, especially in the way we model the membership probabilities, which are allowed to vary across areas. Moreover, it benefits the estimation process in various ways. In the first place, the MoE approach enhance model flexibility while keeping important properties associated with the FH model, such as design consistency and easy interpretation of the predictors. Secondly, our proposal incorporates both the clustering and the small area estimation steps into a unique procedure. In this way, it inherently accounts for the uncertainty involved in the clustering step in the resulting estimator. Eventually, the MoE provides a distinct advantage when prediction has to be carried out for out-of-sample (OOS) areas, as disposing of non-constant group probabilities can be useful in this step.

In this short paper, we present the methodology, overlooking formal proofs, and provide a sketch of the application results. A thorough comparison of various methods using simulation is omitted for brevity.

2 Methodology

Let $\mathcal{U}$ indicates a finite population constituted by N individuals that is partitioned into D small areas $\mathcal{U}_d$. 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; area-specific sample sizes are defined n_d so that $n = \sum_{d=1}^{D_{IS}} n_d$. We denote the WI we are interested in as Y, so that our target of inference is defined as $\theta_d = N_d^{-1} \sum_{j=1}^{N_d} y_{dj}$; where y_{dj} denotes the observed WI of individual j belonging to area d.

The Hàjek type direct estimator we consider, whose definition includes published survey weights w_{dj} is given by

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

In line with the ordinary FH model we specify the following sampling model:

$$\hat{\overline{Y}}_d | \theta_d, S_d^2 \overset{ind}{\sim} \mathcal{N}\left(\theta_d, S_d^2\right), \ d = 1, \ldots, D_{IS}. \tag{2}$$

where D_{IS} is the number of in-sample areas for which $n_d > 0$. Moreover, S_d is the standard error associated to the direct estimate $\hat{\overline{Y}}_d$ and will be treated as known. Actually, we computed it the product of two terms: the standard error under the assumption of simple random sample and a design effect that accounts for all design features (and namely the variable weights and the clustering of observations).

The linking level of the standard FH model that assumes normality is given by:

$$\theta_d | \beta_0, \boldsymbol{\beta}, \sigma_u \overset{ind}{\sim} \mathcal{N}(\beta_0 + \mathbf{x}_d^T \boldsymbol{\beta}, \sigma_u^2), \ \forall d; \tag{3}$$

where $\mathbf{x}_d \in \mathbb{R}^P$ contains the values of P auxiliary variables registered for area d. To complete the Bayesian specification, prior distributions must be specified for both the 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 [4], 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. Lastly, for the regression coefficients, we impose $\beta_p \sim \mathcal{N}(0, 2.5^2), \forall p$, provided that all the covariates are standardised.

We can express the linking level of our FH model extended to include a mixture of experts (FH-MoE) as:

$$\begin{aligned} \theta_d | \beta_{01}, \beta_{02}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_{u1}, \sigma_{u2}, z_d \overset{ind}{\sim} & \ \mathbf{1}(z_d = 1)\mathcal{N}\left(\beta_{01} + \mathbf{x}_d^T \boldsymbol{\beta}_1, \sigma_{u1}^2\right) \\ & + \ \mathbf{1}(z_d = 2)\mathcal{N}\left(\beta_{02} + \mathbf{x}_d^T \boldsymbol{\beta}_2, \sigma_{u2}^2\right), \ \forall d; \end{aligned} \tag{4}$$

where z_d is a dichotomous latent variable that determines the cluster label for area d, assuming values 1 or 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 characterised by distinct coefficients $[\beta_{0k}, \boldsymbol{\beta}_k^T]^T$ and scale σ_{uk}. 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 effects scale σ_v. Additionally, independent Gaussian distributions with a mean of zero and a standard deviation of 2.5 are set for the coefficients

The probability that z_d assumes value 1 is denoted as $\mathbb{P}\left[z_d = 1 | \mathbf{x}_d^*\right] = \pi(\mathbf{x}_d^*)$, where $\mathbf{x}_d^* \in \mathbb{R}^{\tilde{P}}$ is a vector of covariates possibly different to the set used in modelling the location $\mathbf{x}_d$. 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 a logistic regression:

$$\log\left(\frac{\pi(\mathbf{x}_d^*)}{1-\pi(\mathbf{x}_d^*)}\right)\Bigg|\gamma_0,\boldsymbol{\gamma},v_d=\gamma_0+\mathbf{x}_d^{*T}\boldsymbol{\gamma}+v_d,\ \forall d; \tag{5}$$

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

Once we set $\boldsymbol{\Psi}=(\boldsymbol{\Psi}_1^T,\boldsymbol{\Psi}_2^T)$, where $\boldsymbol{\Psi}_k=(\beta_{0k},\boldsymbol{\beta}_k,\sigma_{uk})$, for $k=1,2$, the predictor minimizing a squared loss can be expressed as

$$\begin{aligned}\mathbb{E}\left[\theta_d|\boldsymbol{\Psi},\pi\left(\mathbf{x}_d^*\right),\hat{\overline{Y}}_d\right]&=(1-\tilde{\eta}_d)\hat{\overline{Y}}_d+\pi_P\left(\mathbf{x}_d^*\right)\eta_{d1}(\beta_{01}+\mathbf{x}_d^T\boldsymbol{\beta}_1)\\&\quad+(1-\pi_P\left(\mathbf{x}_d^*\right))(1-\eta_{d2})(\beta_{02}+\mathbf{x}_d^T\boldsymbol{\beta}_2).\end{aligned} \tag{6}$$

where $\tilde{\eta}_d$ is the overall shrinkage factor we can write as

$$\tilde{\eta}_d=\pi_P(\mathbf{x}_d^*)\eta_{d1}+\left[1-\pi_P(\mathbf{x}_d^*)\right]\eta_{d2} \tag{7}$$

and $\eta_{dk}=S_d^2/(\sigma_{uk}^2+S_d^2)$, $k=1,2$ is the expert-specific shrinkage factor.

We remark that we illustrated a mixture with two components, due to features of the motivating application and the tendency for small area models to be parsimonious. However, it is worth mentioning that the results presented here can be readily extended to a model with multiple components if needed.

3 Application

We used DHS survey data for Bangladesh, 2014 wave. A thorough description of these data and the remote sensing (RS) data used as covariates, can be found in [1] and is omitted here for brevity. We only mention that the considered RS covariates include demographics (population density and its Geary index, child to woman and male to female ratios) along with development indicators (night time light radiance, proportion of built and agricultural areas, land use, time health, and other facilities).

These many covariates suffer from quasi-multicollinearity, a problem that we solve using principal components. Specifically, we include the first two principal components as covariates both in the expert network of equation (4) and in the gating network of equation (5), so that $\mathbf{x}_d=\mathbf{x}_d^*$, $\forall d$. Indeed, the inclusion of further principal components does not lead to marked improvements in terms of the popular leave-one-out cross-validation information criterion (LOOIC).

An exploratory hierarchical cluster analysis based on all the upazila-specific covariates was conducted. It confirmed that the optimal number of clusters is two and the identified clusters can be easily labelled as *rural* and *urban*; the first includes peri-urban, rural, and remote areas while the second, smaller in size, encompasses metropolitan and highly urbanised upazilas.

Posterior inference has been conducted using Markov Chain Monte Carlo (MCMC) methods. Specifically, the implementation involved the utilization of

the Hamiltonian Monte Carlo sampling algorithm through the `Stan` language and the `rstan` package. The estimation was carried out using four chains, each comprising 6,000 iterations, where the initial 3,000 iterations were designated as warm-up and subsequently discarded. Concerning the FH-MoE model, label switching problems are avoided as 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 impact θ_d since the linear combination of the two components retains its significance.

The results obtained under the proposed FH-MoE model are compared against the FH model (other comparisons are included in an extended version of this manuscript). FH-MoE is clearly better than FH in terms of LOOIC: 134.3 (28.7) versus 186.6 (35.6). The reason is apparent from table 1, where we can note that the posterior means of regression parameters are different. Namely, the intercepts are widely different and so are the slopes associated to a lesser extent those associated to the second component. Credible intervals have different sizes in the two groups, because of their different size and dispersion.

Table 1. Posterior means and 90% credible intervals related to the basic parameters which rule the models at the linking levels

	FH		FH-MoE			
Comp.	Unique		Cluster 1		Cluster 2	
Par.	Est	90% C.I.	Est	90% C.I.	Est	90% C.I.
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]

More importantly, the posterior of component-specific variance components σ_u^2 are also different, reflecting the different explanatory power of the auxiliary information in the two clusters and namely, the better fit in cluster 1. This implies that mixture-of-experts methodology enables a more pronounced shrinkage in the rural cluster compared to the urban cluster, addressing shortcomings of a standard Fay-Herriot model which would imply uniform shrinkage across areas with differing characteristics. Additionally, the synthetic components to which we shrink differ between clusters due to variations in the relationship between auxiliary variables and the target variable. The different shrinkage processes allowed by FH-MoE with respect to the FH are apparent in Figure 1.

The presence of two regimes distinguished by the FH-MoE leads to notable improvements also in the estimates for OOS areas, as we can exploit the mixture to assign the area for which we do not have observations to its most likely group.

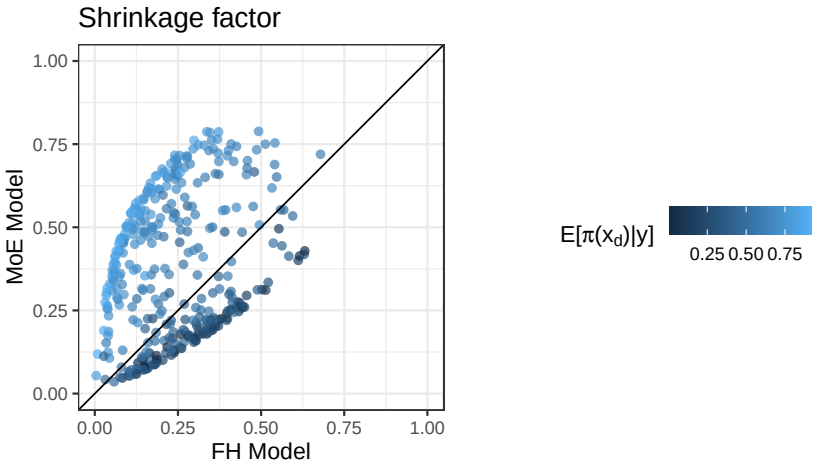

Fig. 1. Shrinkage factors associated to area level estimates computed according to FH and FH-MoE models

References

1. De Nicolò, S., Fabrizi, E., Gardini, A. Extended beta models for poverty mapping. An application integrating survey and remote sensing data in Bangladesh. Technical report, DSS, University of Bologna Forum (2022).
2. Fay R. E., Herriot, R. A. Estimates of income for small places: an application of James-Stein procedures to census data. J Am Stat Assoc, 74, 269-277 (1979).
3. Fruhwirth-Schnatter, Sylvia, Gilles Celeux, and Christian P. Robert. Handbook of mixture analysis. CRC press (2019).
4. Goodrich, B., Gabry, J., Ali, I., Brilleman, S. `rstanarm`: Bayesian applied regression modeling via Stan. R package version 2.1 (2020)
5. Rao, J. N. K., Molina, I. Small area estimation. John Wiley & Sons, (2015).
6. Gershunskaya, J., Savitsky, T. D. 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, (2020)