

1. Introduction

- Model-based methods have been widely used for providing small area estimates (SAE) of population characteristics such as unemployment rate and average income per household. In addition to utilising auxiliary information, typically obtained from Census, being able to estimate variance components reliably has been shown to improve the quality of SAE (e.g., [1, 2]).
- Here, we describe a Bayesian hierarchical modelling approach to deal with heteroscedasticity, where, in practice, the sampling-error variances can differ across small areas. By allowing for unequal sampling variances, the relationship between the response and covariates can be better estimated and hence help improving SAE.

2. SAE models applied to log transformed data

- Distribution of income is typically heavily right-skewed and hence log transformation is often applied to reduce its skewness.
- Let y_{ij} be the income of household j in area i . The following unit level model can be used,

$$\log(y_{ij}) = \alpha + \beta \cdot \bar{\mathbf{X}}_i + u_i + e_{ij} \quad (1)$$

where

- $\bar{\mathbf{X}}_i$ denotes a vector of area-level auxiliary variables;
- $u_i \sim N(0, \sigma_u^2)$ captures area-level variability that is not explained by $\bar{\mathbf{X}}_i$;
- $e_{ij} \sim N(0, \sigma_e^2)$ denotes the sampling errors.

- In this model, the sampling variance, σ_e^2 , is assumed to be constant across areas (hence we label this as the *logConst* model).
- This restriction can be relaxed by letting

$$e_{ij} \sim N(0, \sigma_{e,i}^2)$$

and modelling the **area-specific sampling variances as random effects**, namely $\log(\sigma_{e,i}^2) \sim N(\mu, \sigma^2)$. This is labelled as the *logVary* model.

Area-level SAE

- An estimator of the average household income of area i from the *logConst* model can be written as

$$\hat{Y}_i = \exp \left(\hat{\alpha} + \hat{\beta} \cdot \bar{\mathbf{X}}_i + \frac{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}{2} \right) \quad (2)$$

- The last term, $\frac{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}{2}$, results from the exponential back transformation
→ **the quality of this estimator depends on both the between- and within-area variance estimates.**

- Likewise, an estimator from the *logVary* model is given by

$$\hat{Y}_i = \exp \left(\hat{\alpha} + \hat{\beta} \cdot \bar{\mathbf{X}}_i + \frac{\hat{\sigma}_u^2 + \hat{\sigma}_{e,i}^2}{2} \right) \quad (3)$$

6. Results*

(a) Household comparison (fitted vs. obs.)

- Applied to the original income data, the model allowing for heteroscedastic sampling error variances (i.e., *Vary*) performed the best.
- Model *Vary* achieved 13% reduction in bias compared to the *logConst* model.

(b) MSOA-level SAE comparison

- Both EBLUP and Bayesian fitting produce comparable point estimates;
- Bayesian fitting appears to provide more reliable uncertainty intervals.

* Numeric figures are to be approved for dissemination.

3. SAE models applied to untransformed income data

- An alternative is to fit models exactly the same as *logConst* and *logVary* but to the untransformed income data, y_{ij} (models labelled as *Const* and *Vary*).

Area-level SAE

- For both the *Const* and *Vary* models, the following estimator is used to produce area-level income estimates

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} \cdot \bar{\mathbf{X}}_i$$

→ **SAE from these 2 models do not explicitly involve any variance estimates.**

- But, modelling the sampling variance, $\sigma_{e,i}^2$, hierarchically helps to get a better estimate of the covariate-outcome relationship and hence improve SAE.

4. The Family Resources Survey (FRS) data

- The above 4 models are applied to the FRS data (2008/09) in England and Wales in order to estimate the net household weekly income (after housing costs) at the Middle Super Output Area (MSOA) level.
- The values of $\bar{\mathbf{X}}_i$ are taken to be the means of the corresponding MSOA, in which the household is located.
- All models were fitted in WinBUGS, apart from the *logConst* model where EBLUP was also used.
→ **a comparison of EBLUP and Bayesian fits**
- Due to the survey design, estimation of both the within- and between-area variabilities is done at the Postcode Sector (PCS) level, with respect to which sampling is done at random (≈ 1500 PCS out of a total 8569 have data).
→ In the case of producing MSOA income estimates for the *logVary* model, the estimated PCS-level sampling variances are weighted by population to obtain the MSOA-level sampling variances.

5. Assessing quality of SAE

- We calculate the mean absolute relative bias between the fitted and observed values at the household level:

$$MARB = \sum_{\forall j} \frac{|y_{ij}^{pred} - y_{ij}^{obs}|}{y_{ij}^{obs}}$$

- We compare (a) the posterior means (point estimates) and (b) the corresponding uncertainty intervals of the MSOA-level income estimates (the EBLUP estimates from the *logConst* model are used as the reference):

$$\text{diff}_i = \hat{Y}_i^{model} - \hat{Y}_i^{logConst} \quad (a)$$

$$\text{diff}_i^{IL} = IL_i^{model} - IL_i^{logConst} \quad (b)$$

where IL is the length of the 95% uncertainty interval and $model \in \{logVary, Const, Vary\}$

7. Future directions

- Various variable selection techniques such as reversible jump MCMC (Green, 1995) and the spike and slab method by Ishwaran and Rao (2005) can
 - help improve the selection of “appropriate” covariates;
 - automate the selection procedure.
- Instead of using one single model to produce SAE, Bayesian model averaging approach can be utilised to combine estimates from various competing models. A recently proposed profile regression method (Molitor et al. 2010 and Papatthomas et al. 2011) possesses both features of variable selection and model averaging.