

A COMPARISON OF DIFFERENT METHODS FOR SMALL AREA ESTIMATION

V. Gómez-Rubio^{1*}, N. Best¹ and S. Richardson¹

¹ Imperial College London, U.K. (*v.gomezrubio@imperial.ac.uk)

Introduction

Statistical bureaus often deal with the problem of providing statistics for small areas. The main problem that arises is that data in small areas are sparse and it is difficult to produce reliable estimates. In addition, it is seldom possible to obtain a sample from every small area. For these reasons, to provide an estimate in a small area it is very convenient to use information from other regions. The way information is borrowed from other regions depends on the estimator, but it usually involves the use of regression methods.

Survey design

The design of a survey involves the definitions of the areas to be sampled, the domains (sub-populations) and how the sample is obtained (randomly, etc.).

Direct estimators

This class of estimators are based only on the sample obtained in the survey.

π -estimator

The π -estimator is only based on the sampled values of the target variable and the weights defined in the survey design:

$$\hat{Y}_{i,DIRECT} = \sum_{j \in s_i} w_{ij} y_{ij}$$

Model-based estimators

When only area level data are available, the following model is assumed:

$$Y_i = \beta X_i + e_i$$

where β is the regression coefficient and e_i is the sampling error. At the individual level, the model used is:

$$y_{ij} = \beta x_{ij} + u_i + e_{ij}$$

where x_{ij} are the covariates for individual j in region i , u_i the effect of area i and e_{ij} the individual variability. In matrix form, these models can be written as

$$Y = X\beta + Zu + \varepsilon$$

Z represents the structure of the area random effects and can be used to model different types of effects. The variance of u is denoted by D and that of ε by V . Then, the variance of Y is $\Sigma = ZDZ' + V$.

Synthetic estimator

The synthetic estimator is based on a linear model on some covariates:

$$\hat{Y}_{i,SYNTH(A)} = \hat{\beta} X_i$$

Generalised Regression (GREG) Estimator

This estimator can be defined in different ways depending on how covariates are used, but a convenient form is:

$$\hat{Y}_{i,GREG} = \hat{\beta} X_i + \sum_{j \in s_i} w_{ij} (y_{ij} - x'_{ij} \hat{\beta})$$

where β is obtained by weighted regression. **This estimator combines aggregated and individual data.**

Composite estimator

This estimator is a compromise between the direct and synthetic estimator:

$$\hat{Y}_{i,COMP} = \hat{\gamma}_i \hat{Y}_{i,DIRECT} + (1 - \hat{\gamma}_i) \hat{Y}_{i,SYNTH}$$

where $\hat{\gamma}_i$ is a shrinkage coefficient between 0 and 1.

(E)BLUP estimator

This is similar to the synthetic estimator but an estimate of the area effects is also employed using an unbiased linear predictor:

$$E[u|y] = \hat{u} = DZ'_s \Sigma_s^{-1} (Y_s - X_s \beta) \quad (1)$$

where the s subindex refers to the *sampled areas* or *units*. The BLUP estimator becomes:

$$\hat{Y}_{i,BLUP(A)} = \hat{\beta} X_i + \hat{u}_i \quad (2)$$

Spatial EBLUP

The Spatial EBLUP is based on (1) and (2) assuming a SAR or CAR interaction between the small areas. In the SAR case, the model is

$$Y = \beta X + (I - \rho W)^{-1} u + \varepsilon; \quad W \text{ adjacency matrix}$$

Acknowledgements

This work has been partly supported by BIAS Project, funded by ESRC. V. Gómez Rubio also acknowledges partial support from MTM2004-03290 Project, funded by *Ministerio de Educación y Ciencia* and ERDF.

We would like to thanks N. Salvati for providing the code for simulating the data and computing the EBLUP and SEBLUP estimators.

Bayesian estimation

Using the the model presented before, the posterior mean of the area estimates is exactly the BLUP estimators described before. However, Bayesian models are more flexible and additional spatial and temporal effects can easily be incorporated.

$$\begin{aligned} Y_i | \beta, u_i, \sigma_u^2, \sigma_e^2 &\sim N(X_i \beta + z_i u_i + v_i, \sigma_e^2) \\ u_i | \sigma_u^2 &\sim N(0, \sigma_u^2) \\ v_i | v_{-i}, \sigma_v^2 &\sim CAR(\sigma_v^2) \\ f(\beta) &\propto 1 \\ \sigma_u^2 &\sim Ga^{-1}(a_0, b_0) \\ \sigma_v^2 &\sim Ga^{-1}(a_1, b_1) \end{aligned}$$

Assessing the quality of a model

In order to assess how good is the performance of a given estimator, we have used the Average Empirical Mean Square Error

$$AEMSE = \frac{1}{42} \sum_{i=1}^{42} (\hat{Y}_i - Y_i)^2$$

In a Bayesian setting, the *Deviance Information Criteria* can be used:

$$DIC = \bar{D} + p_D$$

where \bar{D} posterior mean of the deviance and p_D the *effective* number of parameters.

Example

Description of the simulated data

We have considered 42 small areas for which we have simulated the complete population. For each individual, we have simulated a covariate which is used in turn to generate the target variable using a linear regression with coefficient $\beta = 0.21$. The total population is 40236 with an average population of 239.5. The sample size has been chosen at random between 3 and 15. The area effects have been generated so that there is spatial interaction between neighbours. A SAR specification with spatial correlation 0.85.

Results

METHOD	AEMSE	\hat{u}_i	$\hat{\beta}$ (s.d.)	$\hat{\rho}$ (s.d.)	Moran's I \hat{u}_i *
π -estimator	217.211				
GREG	34.914	Yes	0.205 (0.009)		0.418
Synthetic (Area level)	489.641	0	0.206 (0.031)		
Composite (Area level)	416.090	0			
EBLUP (Area level)	154.957	Yes	0.206 (0.031)		0.309
EBLUP (Unit level)	36.775	Yes	0.201 (0.008)		0.408
SEBLUP (Area level)	101.427	Yes	0.212 (0.025)	0.752 (0.135)	0.470

METHOD	AEMSE	\hat{u}_i	$\hat{\beta}$ (s.d.)	DIC (p_D)	Moran's I \hat{u}_i *
Bayesian (Area eff.)	191.681	Yes	0.202 (0.033)	282.364 (39.471)	0.275
Bayesian (+CAR)	167.553	Yes	0.198 (0.025)	282.451 (38.626)	0.356**
Bayesian (CAR)	170.379	Yes	0.191 (0.023)	306.321 (43.659)	0.305

UNIT LEVEL MODELS

Bayesian (Area eff.)	36.934	Yes	0.201 (0.008)	2871.930 (39.918)	0.407
Bayesian (+CAR)	30.325	Yes	0.202 (0.007)	2869.660 (35.535)	0.463**
Bayesian (CAR)	33.162	Yes	0.202 (0.007)	2871.860 (34.187)	0.454

* Moran's I of the actual area effects is 0.499.

** Only the spatial random effects have been used.

Swedish LOUISE Register

The Swedish population register contains different socioeconomic variables at the individual and household level.

Equivalised Income per household

We have used different SAE methods to provide an estimate of the average *Equivalised Income per household* in 1992 using different covariates: # persons in household, # pers. employed in household, and education, age and sex of head of household.

Results

METHOD	AEMSE	\hat{u}_i	Moran's I \hat{u}_i
π -estimator	5494.04		
GREG	4467.08	Yes	0.20
Synthetic (A. level)	7720.83	0	
Composite (A. level)	7415.8	0	
EBLUP (A. level)	1773.4	Yes	0.11
EBLUP (U. level)**	3378.25	Yes	0.31
SEBLUP* (A. level)	1639.56	Yes	0.31
Bayesian (A. eff.)	1816.16	Yes	0.10
Bayesian (+CAR)	1461.8	Yes	0.75***

* $\hat{\rho} = 0.08$ ** Using R pack. nlme *** Only using spatial random effects

Bayesian unit level models

Covariate	Bay. (A.e.)	Bay. (+CAR)
# persons.	-127.13 (6.78)	-126.33 (6.86)
# empl.	271.87 (6.72)	271.15 (6.60)
ed. head	218.53 (9.21)	217.42 (9.21)
age head	13.04 (0.32)	13.02 (0.34)
sex head	93.71 (8.00)	94.49 (8.25)
DIC (p_D)	493739.22 (112.62)	493686.24 (73.92)

References

* Rao JNK (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.

* Salvati N (2004). *Small Area Estimation by Spatial Models: the Spatial Empirical Best Linear Unbiased Prediction (Spatial EBLUP)*. *Technical report*, University of Florence, Department of Statistics.